The International Journal of Language and Cultural
TIJOLAC is Available Online at:
https://www.growingscholar.org/journal/index.php/TIJOLAC
DOI: https://doi.org/10.5281/zenodo.6781613
Vol. 4 No. 1, March 2022, pages: 46~67
Growingscholar Publisher, USA ISSN 2691-4727

Translation Assessors' Perceptions of Assessment Scales Criteria and Its Effect on The Process and Outcome of Their Assessment

Hamed Ghaemi [1], Seyedhamed Sadoughvanini[2]

Article history: Received December, 15, 2021 , Accepted: March 20, 2022 , Published: March, 30, 2022

Keywords

Translation Assessor' perception;

Translation assessment;

Assessment scale;

Analytic scale;

Holistic scale;

Abstract

The present study investigated the Translation Assessor' perceptions of the assessment scales and their subsequent assessment behaviors for two analytic and holistic assessment scales. Hence, nine highly experienced Translation Assessor were asked to verbalize their thoughts while assessment student translations using Translation holistic scale and the analytic scale of ESL Translation Profile. Upon analyzing the think-aloud protocols, four themes emerged. The findings showed that when assessment holistically, the Translation Assessor either referred to the holistic scale components to validate their assessments (validation) or had a pre-evaluation reading to have more reliable assessment (dominancy). In analytic assessment, on the other hand, the Translation Assessor used a pre-evaluation scale reading in order to keep the components and their criteria to memory to evaluate the text more accurately (dominancy) or regularly moved between the text and the scale components to assign a score (oscillation). Furthermore, the results of a Wilcoxon signed-ranks test showed that when using the holistic and analytic assessment scales, the Translation Assessor assigned significantly different scores to the texts. On the whole, the results showed that the way the Translation Assessor perceived the scale components will affect their judgements of the texts. The study provides several implications for rater training programs and EFL translation assessment.

[1] *Bahar Institute of Higher Education, Mashhad, Iran. Email: Ghaemiacademy@gmail.com (https://www.scopus.com/authid/detail.uri?authorId=36720430300)*
[2] *Bahar Institute of Higher Education, Mashhad, Iran*

1. Introduction

The outcomes of the translation assessment derive from an interaction among many elements of assessment including rater, test-taker, scale, scoring, and practice. Therefore, the quality of assessments is a critical reflection in the interpretation of outcomes (Hamp-Lyons, 2007), and hence different interactions regarding all these aspects of assessment can lead to different outcomes. In translation assessment, the impact of assessment scale on the assessment process (Barkaoui, 2010a), complexity of assessment criteria (Joe, Harmes, & Hickerson, 2011), and Translation Assessor' perception can induce the lowest inter-rater reliability. On the other hand, a translation has lots of features that are the basis of assessment; therefore, identifying the text components and bringing them in a conformity with the scale's components is a demanding job. To increase the reliability of the assessment, Translation Assessor should not misconceive the assessment components which might lead to a decrease in the consistency of the assessment performance. In other words, different factors contribute to inconsistency in the assessment process, which Translation Assessor' perceptions and justifications of scoring decisions are more highlighted. According to McNamara (1996), there are many factors involved in performance assessment in which assessment scale can play a permanent for its role in increasing the quality level of judgement. To thrive in this process, Translation Assessor must adopt and use an appropriate assessment scale. They should sift through them so that they can choose one regarding their purpose of assessment translation. Generally, the assessment is a formidable job entailing a rater' sufficient knowledge, judgment, and the skill of interpretation. The significant role of the assessment scales in translation performance increases the degree of Translation Assessor' interpretive and judgmental accuracy. One way of achieving accuracy is that Translation Assessor concentrate precisely on the assessment scale during the assessment process. In order to accomplish this, they should perceive the components of the assessment scale precisely so that they can make a fair judgement. The evaluation of translation is a subjective process, and human Translation Assessor predominantly rate the texts using assessment scales which might affect the assessment practice. In fact, without having a clear picture of what happens in the Translation Assessor' cognitive process, it is impossible to claim whether their evaluations and subsequently the assessment outcomes are at all fair or not (Sakyi, 2000; Cumming, 2002). As a result, this shortage puts the scoring validity and consequential validity at risk (Weir, 2005). It is not expected that all Translation Assessor perceive the assessment scales' components in a consistent and similar way. It is also inconceivable to gain the uniform results from all the Translation Assessor.

Most of the previous scale effect studies were limited to comparing different types of scales (e.g. Bacha, 2001; Barkaoui, 2010, 2011; Knoch, 2007). Furthermore, despite doing numerous researches on rater behavior in the framework of performance assessment, few researchers have addressed the problem of rater's perception in the process of assessment in EFL contexts (Eckes, 2006). As long as assessment and assessment are performed by humans, they are inevitably influenced by personal beliefs, ideas, and preferences (Eckes, 2008). The components of the assessment scale, the wording of the rubrics, the impressions brought with Translation Assessor, and the potential interaction of these elements can affect Translation Assessor' perceptions, and therefore, the scores they assign (Goodwin, 2016). Due to the vagueness and lack of explicitness in the application of assessment scales' components, Translation Assessor may not have an explicit perception during the assessment process, and it may affect the outcome of assessment.

As a result, to gain a clear picture of the Translation Assessor' perceptions of assessment scales and their effects on the assessment process and outcome, in this study the interaction between rater's perceptions and assessment scales, were investigated. The main concern pursued in this research is considering the Translation Assessor' perceptions of assessment components regarding two different assessment scales which might be significant in a number of ways: First, because there is a possibility

that Translation Assessor do not have the explicit perceptions of assessment components, the misperceptions may decrease the reliability of the assessment. If the Translation Assessor perceive the assessment components precisely, this will help teachers and Translation Assessor during the assessment process to get a better outcome. Second, this explicitness also helps the teachers who evaluate students' translations holistically and analytically through focusing on textual features in the translation class which lead to the development of students' translation skill. In other words, a clear perception of the assessment scales' components can also help the Translation Assessor focus on other translation features. Third, it seems that lack of explicit and precise perception of assessment rubric leads to unfair scoring. If the error of measurement reduces through consistency of assessment, accuracy and fairness of Translation Assessor will increase. Fourth, according to Russikoff (1995), when Translation Assessor rate ESL texts holistically, they pay more attention to language use despite the robust content and organization. Fifth, according to Weigle (2002), in a holistic assessment scale, descriptions of features are combined and thus make the score challenging to interpret; Therefore, the findings of this study can help the Translation Assessor have a robust interpretation of the test scores that further increase the validity of the translation assessment.

2. Literature review

Translation assessment is considered as one of the difficult tasks in language assessment (Coombe, 2010). In translation performance assessment, the performance process is highlighted, observed, and judged according to an assessment scale. In spite of many advantages drawn from performance assessment including more fairness, accuracy, and objectivity, it might not be void of problems. It is evident that the process is time-consuming and challenging. Furthermore, Speck and Jones (1998) assigned issues such as "inter-grader reliability, single-grader consistency, and ultimate accountability" (p.17).

The characteristics of Translation Assessor in translation assessment are the idiosyncrasy that can be sought in the rater's cognitive, efficacy, and perception; these factors are affected by Translation Assessor' appropriate education, training, and prior experience (Rahayu, 2020; Khodi, 2021). To prevent assigning inconsistent scores, concentassessment properly and exclusively on evaluation of examinee' translation performance during assessment process, and applying appropriate assessment procedures, as well as a training course can be followed by teachers or Translation Assessor. However, the knowledge of translation assessment and the efficacy of Translation Assessor in scoring accuracy are the basic necessities, the correct perception in assessment process can ascertain the how fair and effective they rate a paper (Meissel, Meyer, Yao, & Rubie-Davies, 2017). This confirms different Translation Assessor' perceptions as a source of inconsistency in the assessment process (Davidson, Howell, & Hoekema, 2000).

Perceptions have been identified as idiosyncratic and subjective notions, views, and evaluations of someone's performance which significantly contribute to second language teaching and learning because they form the base of teachers', test-takers', and Translation Assessor' views, judgments, and interpretations (Brown, 2009). Ono, Yamanishi, and Hijikata (2019) defined Translation Assessor' perceptions as "important aspect of Translation Assessor' scoring behaviors and views of (in) effective performance, as well as the challenges they face while scoring" (p.70). Barkaoui (2007) investigated the effects of assessment scales on Translation Assessor' perceptions, outcomes, and scoring processes. Four Translation Assessor rated 24 translations silently adopting a holistic scale and then two subsets of four translations applying a holistic scale while thinking aloud. The findings showed that in contrast to holistic scores, which were more reliable, there were inconsistencies in multi-trait scores which might be because of not training the Translation Assessor. Among all aspects, the content and organization were of higher reliability than style, grammar, and mechanics. In an explanatory study on integrated written tasks, Ono et al. (2019) compared teacher-Translation

Assessor' using holistic and analytic scales for reliability and validity to unravel their perceptions of assessment. The findings illustrated that Translation Assessor encountered various challenges while scoring. According to Panadero (2020), as analytic scales focus on the components rather than the whole, the assessment process may undergo the fragmentation. In another study, Eckes (2008) ascertained the extent of differentiation among the Translation Assessor' perceptions of criterion importance and examined the performance of the importance assessment scale. Findings revealed that Translation Assessor acted differently regarding the significance of the various criteria. Besides Translation Assessor' different perceptions, to some degree, rater background variables led to scoring profile differences.

Humphry and Heldsinger (2019) investigated perceptions of assessment criteria related to identifying translation performance levels. Anchoring direct pairs of translations, Translation Assessor determined the well-organized and proficient translation according to an analytic assessment scale consisting of ten ordered criteria and a holistic scale. The results illustrated that Translation Assessor considered some features more prominence over others, such as ideas and text structure. In a seminal work, Hijikata-Someya, Ono, and Yamanishi (2015) examined to what extent the Translation Assessor' perceptions of Native-English-speaking and Nonnative-English teacher-Translation Assessor differ when assessment a translation using the holistic scale. Findings illustrated that the former group of Translation Assessor perceived the evaluation of two components namely content and language use as difficult; In contrast, the non-native teacher-Translation Assessor considered assessment challenging as associated with vocabulary. In a mixed-method research, Wang et al. (2017) investigated Translation Assessor' perceptions and judgments assigned scores to integrated translations (reading to write task). Through analyses of Translation Assessor' perceptions, researchers figured out that there was an inconsistency in Translation Assessor' perceptions towards particular aspects of translations.

The current study

The analysis of Translation Assessor' perceptions can reveal the critical features of the scales components and also identify how the Translation Assessor behaved when facing difficulties with the assessment components. In this study, the aim was not to compare but to identify and categorize the Translation Assessor' perceptions regarding the components of the assessment scales.

Translation Assessor always face various challenges to score accurately which may originate from different perceptions towards the scale components. This study sought for a better understanding of the interaction between the rater and the scales components and how it might affect the process and outcome of their assessment. Hence, to achieve these goals, the following research questions were posed in this study:

1. How do Translation Assessor perceive the assessment components when using holistic and analytic assessment scales?

2. Do different Translation Assessor' perceptions affect the assessment outcome and process using holistic and analytic assessment scales?

3. Methodology

In this study, a triangulation mixed-methods design was adopted. Both qualitative and quantitative data was collected to identify the Translation Assessor' perceptions regarding the assessment scales components. The first phase was qualitative which focused on the data obtained

from think- aloud protocols (TAPs). Because translation assessment process is not easy to identify, then conducting a qualitative research is appropriate for studies that are not easily measured (Creswell & Poth, 2018). A qualitative approach could assist the researcher in effectively interpreting the depth, variety, and complexity of Translation Assessor' thought processes. In the second phase, the outcomes of the assessments were quantitatively analyzed in order to find out if assessment using different assessment scales would affect the score the Translation Assessor assigned.

The Translation Assessor who participated in this study were nine experienced Iranian EFL teachers who were selected using convenience sampling from two Iranian state universities including Persian Gulf University (PGU) and Salman Farsi University of Kazerun (SFKU). These Translation Assessor were teaching the undergraduate courses of advanced translation and translation.

They were experienced translation teachers with at least five years of teaching and assessing translation experience in the university. The age range of the participants was between 35 and 59 years old. The research participants were chosen from both female and male teachers (five males and four females). Four out of five male Translation Assessor had MA degrees, and the remaining one had PhD. In addition, three out of four female Translation Assessor hold PhD. and the remaining one had MA. Also, only three out of nine Translation Assessor had participated in a workshop related to assessment translation assessment. All the 9 Translation Assessor agreed to contribute to the study. They were also assured that their data would be confidential and would be used only for research purposes.

Instruments

In order to collect data for the present study, six instruments were used including translation tasks, two assessment scales including TRANSLATION Holistic Scale and ESL Translation Profile (Jacobs, et al., 1981), training manual, demographic information sheet, and think-aloud protocols (TAPs). The following briefly explains each.

Translation task

The translation sample used in this study were randomly selected from a body of thirty translations written under exam conditions by TEFL undergraduate students. As the final exam of the translation course, the students were asked to write a translation (about 160 words) on the topic "how do you recover emotionally from the disasters". The translation was used for two rounds of assessment and scoring holistically and analytically. The Translation Assessor were next asked to rate the translations based on the holistic and analytic assessment scales they were provided with.

Translation Holistic Scale

Translation assessment scale as the holistic scale was used in the study. In this scale, a text is judged by Translation Assessor as a whole while parts of the translation are not separated and identified. The text is given a score from 1-9. ESL Translation Profile (Jacobs, et al., 1981) ESL Translation Profile (Jacobs et al., 1981) as an analytic assessment scale was used in this study. The reason behind selecting this instrument was Sasaki and Hirose' (2004) who claimed that Jacobs et al.'s (1981) ESL translation profile was a commonly-used scale and one of the most preliminary well designed rubrics to assess L2 translation though not without its pitfalls. The rubric consists of five main components of a written text: content and organization which refer to the prepositional content, unity, cohesion, and coherence of the text; vocabulary refers to the range, lexical diversity, word choice, and word form mastery; language use focuses on grammatical points, accuracy, and syntax; and last but not the least is mechanics which refers to paragraphing, capitalization and generally

superficial aspects of a text. Each component is rated according to four levels from very poor to very excellent. The scores for each level in each component differ.

Demographic information sheet

In order to obtain the background information of the Translation Assessor, they were asked them to fill in a demographic information sheet containing a personal profile, current professional status, educational history, and the experience of assessing and teaching translation.

Training manual

Training manual consisted of an instruction sheet which the Translation Assessor received in order to guide them to conduct TAPs in a consistent way. In addition, along with this manual, the researchers sent a recorded message in which they explained the think-aloud procedure precisely and provided some examples in order to clarify and adequately convey information about how to perform TAPs and also apply holistic and analytic assessment scales.

Think-aloud protocols (TAPs)

In general, to analyze translation assessment processes and map out the Translation Assessor' mental processes when assessment, the participants performed TAPs. The procedure involved verbal expression of their mental processes. For this, the Translation Assessor were trained to verbalize their thoughts while assessment a translation. The Translation Assessor' verbalizations were recorded to be analyzed next.

Data Collection Procedure

Early in the study, a group of qualified EFL translation instructors were contacted to participate in this study. Due to the difficult situation caused by the COVID-19 pandemic, many teachers were reluctant to participate in face-to-face sessions. Therefore, a package including two assessment scales, a training manual which consists of an instruction of TAP along with a voice recorder, translation texts, and demographic information sheet was emailed to the teacher Translation Assessor. Data collection proceeded in four stages: first, a personal demographic information was given to the Translation Assessor to obtain the Translation Assessor' information such as age, gender, experience of teaching and assessment translation. Second, a training manual was given to the Translation Assessor in order to make them familiar with the TAP procedure. Third, they were asked to play the recorded voice which was attached to the manual to clarify how to do TAPs and also rate with the two scales as well. It is also worth mentioning that the Translation Assessor were free to use either mother tongue (Persian) or English in performing TAPs. Fourth, the TAPs were carried out in order to see into the Translation Assessor' minds and investigate their thought processes during the assessment process. Therefore, the Translation Assessor were asked to verbalize their thoughts as they were assessment the translation. They used two different assessment scales.

After assessment using a particular scale finished, the easy received the score and immediately the next scale started. The Translation Assessor started recording their voice upon beginning assessment the translation. On average, the process took about 45 minutes for each rater. Then, the recorded TAPs were transcribed in order to codify their utterances and identify some patterns among them.

Data analysis

In order to analyze the Translation Assessor' perceptions of the scales components during the assessment process, the assessment behaviors of nine Translation Assessor obtained from TAPs were transcribed and then codified to seek for the general patterns emerging from the data. Qualitative content analysis was also used to analyze data. The analysis proceeded in four stages; In order to familiarize with the data, the researcher read through the transcripts before breaking down them into segments and meaningful units. Coloring highlighters (pens) were used to recognize and understand the difference among codes. The identification of each meaningful unit labeled with a code, was the second stage. However, due to the nature of inductive coding, the codes changed as the study made progresses and coding process repeated several times.

Additionally, it was checked whether all aspects of the content of transcriptions were covered regarding the purpose of the study. The third stage was categorization. In this stage, the codes were condensed meaning that fewer words were used without missing content of the meaningful unit which led to the identification some themes. The fourth stage was compilation, when the emerged themes were further refined. Moreover, to answer the second research question, using SPSS program Wilcoxon Signed Ranked Test was used.

4. Results and Discussion

Investigating the first research question

To answer the first research question about general behaviors of the Translation Assessor and their interactions, the holistic and analytic scales are presented. Next, a focused analysis will reveal how the Translation Assessor perceive the assessment components when using holistic and analytic assessment scales. It is worth mentioning that out of the nine Translation Assessor, three misfit Translation Assessor were identified, meaning that they did not attend to the scales' components for assessment the texts.

General assessment behaviors

The results showed that the Translation Assessor referred or consulted the assessment components of the scales at different times and with different frequencies. In this study, the assessment processes of 9 Translation Assessor first regarding the holistic scale and then analytic scale were analyzed. Six out of nine Translation Assessor applied holistic and analytic assessment scales' components. Table 1 below shows the interactions of the six Translation Assessor in the study. As shown, in the left side of the table, the codes based on the Translation Assessor' behaviors have been presented, and in the right side, the Translation Assessor' excerpts have been provided.

Table 1: A sample of the Translation Assessor' behaviors interacting with the holistic components

Behaviors	Excerpts
Reading the band score (R9)	*To evaluate holistically, actually the definition of expert user is: test taker has fully operational command of the language. Their use of English inappropriate, accurate ….*

Reading the band scores down up in turn (R2)	*Well! I read the band scores from bottom I start with band score 1*
Skipping reading the components of first two band scores (R3)	*let me see so if I want to grade this holistically let me see umm, she was not an expert user actually she was not very good user actually*
Referring to scale to read band scores' components (R6)	*Now I'm going to look at the scale and read the band descriptors of that score. from holistic point of view… The test taker has a partial command of the language and copes with overall meaning….*
Insisting on her initial impression of the text (R4)	*Did not read other band scores' components because based on what in mind is the translation falls in needless to read others*

As illustrated in Table 1, the Translation Assessor displayed a wide range of behaviors at different stages of the assessment process. In terms of interaction with the assessment scale' components, the Translation Assessor mostly behaved differently. Some behaviors of the Translation Assessor regarding the analytic scale components are shown in Table 2.

Table 2: sample of the Translation Assessor' behaviors interacting with the analytic components

Behaviors	Excerpts
Pre-evaluating the scale components (R9)	*First of all, I need to read the assessment scale, reviewing Its components, well… content, its criteria, …. Organization, um…ok, language use, four levels for each component*
No attention to one component when reading the text (R6)	*Um, you know I did not look at the text carefully to see whether the writer observed punctuation, capitalization…*
Assigning score to the component (content) after reading the criteria (R1)	*I think content was um translation falls between 21 to 17 because (long pause reading the criteria) the development of ideas is inadequate .. development so 21 to 17*
Referring to the content component while assessing and reading organization (R2)	*not fluent, ideas confused or disconnected…. Well! the real problem of the text was related to its content not confusing ideas*

Pause for reading the component criteria (R1)	*I think in terms of uhmm (long pause) grammatical use*

Translation Assessor' perception of holistic scale

Upon the analysis of the Translation Assessor' interactions with the holistic scale components, two major themes of validation and dominancy emerged. With regard to the validation theme, the Translation Assessor rated the text impressionistically and then assigned the score. Next, to validate their scoring, they referred to the scale and read the band descriptors of the scores given. Reading of the band descriptor persuaded them to revise the score. In dominancy theme, the Translation Assessor scanned the scale before beginning to rate so that they could review the band descriptors and develop a master-plan of their evaluation. In what follows, the stages the Translation Assessor went through in their assessment processes for both themes are delineated (Table 3).

Table 3 Translation Assessor' interactions with components of holistic scale

Themes	
Validation	1. Self-monitoring focus 2. Assigning score 3. Referring to scale to read the band descriptor 4. Revision 5. Verification
Dominancy	1. Pre-evaluation scale reading 2. Detailed assessment 3. Consulting scale to reread band descriptors 4. Assigning score 5. Verification

Validation

According to Table 3 above, two Translation Assessor (R1and R6) applied this theme in their assessment. Both Translation Assessor explicitly mentioned that their assessment would be based on the holistic scale. For example, rater1 mentioned that, 'scoring the translation based on the holistic method, ok! I will read the text first and then I will have my overall assessment of the passage'.

Initially, both Translation Assessor articulated a general impression of the translation. Their meticulous evaluation based on self-monitoring focus assured them to assign score before reading the scale's band descriptors. Obviously, the score given was in accord with their initial impression. To validate their scoring, they referred to scale to read the components of the score given. Reading the band score's components made them to revise the score subsequently giving evidence from the text. It can be seen here:

R6:' From a holistic point of view, I cannot assign a score higher than 5, yes 5 is good. All right and in scale (pause)the test taker has a partial command of the language and copes with overall meaning in most situations (band score), possibly 4: basic competence is limited to familiar situations... Yes. I assigned 4, because the content was irrelevant to the topic, it also was not well-organized, but the grammar was satisfying.'

As it is shown below, reading the band descriptor of the assigned score led to the revision of that score and rater selected a different score:

R1: 'I think (pause)I think hum the score for this text falls in um (pause) 6 band which says the test taker has an effective command of the language despite some inaccuracies, inappropriate usage and misunderstandings. He/she can use and understand fairly complex language, particularly in familiar situations. No! I think I give it 7, because the vocabulary the writer has used is appropriate and… pause…. I give it 7.'

The above quotes illustrate how by using the scale's components, the Translation Assessor made or revised their scoring decisions. After revising the score, the Translation Assessor provided evidence from the text to justify their assigned score. For instance, R6 reminded some of the translation errors based on the components in the band descriptor 4 which says because she did not address the topic, did not use complex structures, also kept meandering on.

These Translation Assessor did not read other band scores' components. They only read the band descriptor which they thought was appropriate for the translation and after reading the components, selected another one which was only one band higher or lower the initial score. It seems that the Translation Assessor were already familiar with the band scores' components and were able to make an approximate estimate of the score.

Dominancy

Here, the Translation Assessor scanned the scale before starting the evaluation process. The reason was to get a general picture of the whole scale and adequately dominate the scale band descriptors during the assessment process. For example, Translation Assessor 2 and 3 clearly stated that they had a pre- evaluation reading:

R2 'I prefer to read this scale before correcting the translation; however, you explained it explicitly, just for reviewing, …um nine bands, you now it is general focusing on definite guidelines, expert users say…'

R3:' I took a look at the scale, scanned it, to me now I can have a good evaluation.'
After assessment the text accurately, the Translation Assessor consulted the scale to assign score. The band scores and their components were reviewed so that this whole reading of band scores could help them reach the appropriate score. Re-reading the band descriptors, analyzing them and giving evidence from the text led them to pay attention to different aspects of translation and employ a larger and more varied number of the translation-assessment criteria.

R4:' Based on the band scores components, the translation should contain accuracy, appropriate usage without misunderstanding, convey a communication. Actually, the writer has somehow tried to observe these but generally about the content and organization she should work more on these aspects…'

Interestingly, R2, R3, and R9 who had a pre-evaluation of the scale read almost all of the band descriptors. Among all these four Translation Assessor, only R4 did not read the whole band scores' components. She returned to scale to only read one band descriptor which was in accord with her initial impression. As evidence, the way rater 3 interacted with the text and the scale comes below:

R3: '…she was not an expert user actually she was not very good user actually good user: The test taker has operational command of the language, though with occasional inaccuracies, inappropriate usage and misunderstandings (long pause) well I saw some complex language m most of sentences were clear. Competent user: The test taker has an effective command of the language despite some inaccuracies, inappropriate usage and misunderstandings, the content and organization were not ok, no, modest user, let me check this one modest user: The test taker has a partial command of the language and copes with overall meaning in most situations, although is likely to make many mistakes. He/she should be able to handle basic communication in own field. of course, I think that the vocabulary she used were in advanced level (pause)... but little knowledge on content and organization of translation….'

Overall, for both themes, the Translation Assessor focused more on strengths of the translation using the scale components. Also, content and organization among all the scale components received more attention. Moreover, when the Translation Assessor encountered the difficulties, they used resolution strategies such as anchoring the holistic scale with other scales or comparing two band descriptors or using analytic assessment scale criteria while adopting the holistic assessment. Here is an example of anchoring two band descriptors by R3:' the writer has lots of problems related to the complexity. This band deals with the use of complex language, while the previous one focuses on communication.'

Translation Assessor' perception of analytic scale

The analysis of the Translation Assessor' TAPs showed that in the analytic assessments, the Translation Assessor frequently channeled their attention to details. All the Translation Assessor expressed their initial impression of the translation through looking at the layout, cross outs, following assessment the translation precisely. However, their assessment and scoring processes were based on the scale components. In addition, their interactions and behaviors were fundamentally different. Generally, Translation Assessor who interacted with the scale components applied two themes of dominancy and oscillation in their assessment processes. In the dominancy theme, the Translation Assessor had a pre-evaluation scale reading in order to keep the components and their criteria to memory; therefore, they could evaluate the text more accurately. For the oscillation theme, the Translation Assessor regularly moved between the text and the scale components. In this section, both themes are elaborated as below:

Table 5: Translation Assessor' interactions with components of analytic scale

Themes	
Dominancy	1. Pre-evaluation scale reading 2. Appraising the text 3. Consulting the scale 4. Re-reading components criteria 5. Assigning score
Oscillation	1. Continuous move between the text and the scale 2. Reading component criteria 3. Assigning score

Dominancy

Regarding the dominancy, the Translation Assessor (R1, R2, R3, R9,) scanned the analytic assessment scale (Jacobs, et. al., 1981) prior to their evaluation in order to include the components into their assessment and take into account the multi-faceted nature of translation aspects. As an example, rater3 emphasized, 'I have not used this scale for a long time, it's better to take a look at it, five components, yes, content, um, so, knowledgeable, substantive, development of thesis…'

Additionally, this pre-evaluation scale reading was done because they wanted to review the components and the criteria of each component. Then all four Translation Assessor precisely rated the text and consulted the scale to articulate score decision according to the components: R1:' ok this is the paragraph, let me read for the details and the grammatical (pause) lexical, organization. (pause) and the rest of the components of the analytical scale…. '

As shown below, Rater 2 read the criteria of all levels in each component reversely in order to reach an appropriate level, interpreted, translated them into Persian, and commented on the mistakes of the text according to those criteria so that he could select one of the levels as the final decision.

R2: 'In language use, again, at first, I read the last level which is very poor: virtually no master of sentence, construction dominated by errors, no, the communication can be seen in the translation some complex sentences for example the second line of first paragraph, did not have lots of errors, fair to poor: major problems in simple and complex construction, frequent errors of negation, agreement, it's not acceptable, the writer has used passive form very well, good to avenge: effective but simple construction, exactly she used simple structures but correctly, several errors of agreement yea, as I told before the problem of dangling which is related to agreement, I select 21-18 for language use.'

On the other hand, rater 3 read only two levels of each component:
'First the content, fair to poor: limited knowledge of the subject, little substance (pause), let me check the last one 16 to 13 should be very poor, does not know knowledge of the subject, no, I think about the content, the third one is ok, so 21 to 17.'

Rater1 read only the criteria of one level for each component in agreement with his general impression which indicated that he was impartially certain about that score because after reading the criteria he did not revise the score:

R1:' the content (pause) I think content was um translation falls between 21 and 17 because (long pause reading the criteria) the development of ideas is inadequate ... development so 21 to 17. In terms of organization it can fall fair to poor: non-fluent, ideas confused or disconnected (pause) although the paragraphs are not well-developed, the way has written sentences ...(pause) and ideas are linked, it falls 17 to 14.'

Rater 9 read the criteria of first level (Excellent to very good) in each component as the definition of that component and then selected one level out of three regarding her general impression to read its criteria. When the rater wanted to assign score for one of the components (mechanics), she was dubious about that component which resulted in consulting the text and assessment the component again.

R9: 'about mechanics excellent to very good (pause) well actually I'm a bit doubtful about this component, when I was reading the text, I didn't pay attention to this component, looking back the text, some cross outs, hand translation…'

Oscillation

Following the oscillation theme, Translation Assessor (R4 and R6) moved continuously between the text and the scale. The Translation Assessor had a separate evaluation for each component in detail and then read the criteria of that component to articulate score. As an example (the words in bold are the component' criteria).

R4: 'In terms of organization, what I see here is that excellent to very good: fluent expression, actually the first paragraph has been written partly well , some errors are seen here for example the ideas have not developed there I consider fair to poor : non fluent, ideas disconnected, there is no harmony between the thesis statement….ideas are clearly stated, um no, it started with Natural catastrophes lead to higher costs of living but did not supported it, well, good to average : somewhat choppy, loosely organized… yes as I see ideas are not connected, limited support, logical but incomplete…, lots of irrelevant details, good to average is reasonable…..in terms of vocabulary….'

It seemed this style of assessment components could not help them in decision-making process since they had simultaneous conflicting reaction towards each component. In other words, they showed ambivalence in assigning score even after several times reading the criteria and giving evidence from the text. Therefore, the scale components did not provide adequate and explicit information for the various aspects of the text and might block the interpretations of the Translation Assessor.

R6: 'In terms of vocabulary, I choose very poor because completely describes the translation: Essentially translation, little knowledge of English vocabulary, idioms, word form, actually she has little knowledge of this topic, no word related to topic, not idiom, of course now I'm seeing that in the last part of this level it has said Not enough to evaluate, so let me read another one fair to poor: Limited range, frequent errors of word form, idiom form, choice, usage, meaning confused, um, I think it's ok, I assign this score to this component… um let me return to text I see she has used look for in wrong way, lead to, um , cause or for example damaged buildings leave no place for people to live without worrying their lives.. No, I select the previous one very poor because look for was translated.'

Similar to Rater 9 in the dominancy Theme, R6 forgot to pay much attention to mechanics component and consulted the translation to scan the text and look for the mechanics errors:

R6: ' Uhmm, mechanics, very poor: No mastery of conventions, dominated by errors of spelling, punctuation, capitalization you know I did not look at the text carefully to see whether the writer observed punctuation, capitalization(pause) I see that there is a comma after however, poor handtranslation, Frequent errors of spelling, punctuation, capitalization, paragraphing, poor handtranslation, right, 3 is satisfying.'

In fact, although the Translation Assessor applied the same scale to justify their scores, they adopted different approaches in their evaluation. For instance, some Translation Assessor disregarded a component in assessment process and when they wanted to mark in scoring process were obliged to consult the text in order to re-rate that component. While R6 and R9 returned the text to assess the mechanics, R6 usually resorted to the strategy of comparing two levels of each component to settle

down her problem about the vagueness of scale wording. Moreover, more detailed and cautious evaluation involved iterative assessment and scoring behaviors with too much of a cognitive load. That is why when Translation Assessor were not able to reach a sound decision on yielding score for each component tend to interpret the criteria meticulously or compare the criteria of two levels or two scales.

According to the dominancy, pre-assessment scan of the scale helped Translation Assessor select a level out of four more easily and master the scale' components. This group of Translation Assessor, in their consultation of the scale, tried to re-read most of the levels' criteria. On the other hand, in oscillation, the Translation Assessor were doubtful in selecting one level and assigning score for that component.

Investigating the second research question

The second research question aimed to examine whether the Translation Assessor' perceptions of the two scale components influence the assessment outcome. To answer this question, Wilcoxson Signed Ranked Test was used to examine the possible differences between the scores assigned by the analytic and holistic scales.

Table 6: Descriptive statistics of holistic and analytic scales scores in this study

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Holistic	9	5.44	1.424	4	7
Analytic	9	62.11	13.606	40	76

The results of the descriptive statistics on the scoring scales indicated the minimum and maximum of scores in two different assessments and standard deviation that showed the spread of the data. In holistic scale, scores were computed from 9 while in analytic scale, they were calculated from the total score of 100.

Table 7: Wilcoxon Signed Ranks Test for the two Scales

Test Statistics

	Analytic - Holistic
Z	-2.666[b]
Asymp. Sig. (2-tailed)	.008

As the table above shows, the significance level is less than 0.05 ($p<0.05$) which indicates that there was a difference between the Translation Assessor' scores assigned holistically and analytically. In other words, there was a significant difference between the mean of the two sets of scores meaning that the Translation Assessor assigned significantly different scores to the text using the holistic and analytic assessment scales. Therefore, the second hypothesis was rejected. In other words, having different perceptions on the assessment scales' components, and the way the Translation Assessor interacted with them affected the assessment outcome assigned.

Discussion

The analysis of TAPs showed that the Translation Assessor adopted different assessment behaviors in two kinds of their evaluations. Regarding the first research question, the most striking results to emerge from the data are that the Translation Assessor referred to the scales' components prior to their scoring, after looking for mistakes, before and after assigning the score holistically and after scoring to each component of the analytic scale separately. It was also found that three Translation Assessor followed dominancy in their both holistic and analytic assessments. They adhered to the guidelines of scales and preferred to rate the texts considering the whole components of the scales.

All in all, about 50 assessment behaviors were coded in the present study. The Translation Assessor in the analytic assessment mode showed a wider range of these assessment interactions and complicated thought processes probably because they had not used the scale increasingly or because the scale components did not cover all possibilities, Translation Assessor tended to formulate various strategies to help them deal with problematic dimensions of the assessment process (Lumley, 2002) such as ambivalence, disregard of components and complexity of the scale wording.

The Translation Assessor in the analytic assessment considered the scale components in two themes two times. In the dominancy theme, once the Translation Assessor had a pre-evaluation scale reading had an interaction with the scale' components and once again when they consulted the scale after precisely assessment the translation so as to read the components' criteria. In the oscillation theme, once the Translation Assessor referred to the scale to figure out which component they had to consider and also scan the criteria and once again when they consulted to read the criteria precisely and make comments.

Simple application of scales does not certify an effective evaluation, and an ill-designed assessment scale might even block understanding of criteria included in the scale. On the other hand, as Panadero and Jonsson (2020) claimed, assessment scales are scoring guidelines that contain described components at different levels of standards. As mentioned before, for both themes the Translation Assessor had some difficulties with the scale components which could have been derived from lack of familiarity with the application of the scale or complicated nature of the scale. These strategies included anchoring, translation, much scrutiny of assessment components, reading reversely, and self-knowledge-based inference of the components.

Moreover, the content of analytic scale increases the working memory of the Translation Assessor (Hirai & Koizumi, 2013) because of having a wide range of criteria. Thanks to this, it might be difficult for Translation Assessor to analyze and interpret the scale components and through adopting those strategies, they lessened the cognitive load. On the other hand, findings of Jeong (2019) showed that the scale design has a greater effect on the Translation Assessor. Thus, designing assessment scales and the assessment criteria should provide an explicit and reliable foundation for scoring judgments, over and above distinguishing translation performance levels (Weigle, 2002). This type of assessment scale (Jacobs et al., 1981) used in this study has been criticized for the vagueness of the components' criteria. Turner and Upshur (2002) criticized the criteria for being irrelevant to the test task and its context and the criteria are improperly grouped at descriptor levels.

An example can further clarify this observation. Two of the Translation Assessor expressed their ambivalence in assigning score to all analytic scale components. According to the observations, Translation Assessor felt more confusion in their assessments which implied that they got it quite demanding to clearly selectone level. This finding is in line with Jeong (2019) which concluded that Translation Assessor had to consider multiple areas in the analytic assessment process and experienced more hesitation and assessment conflict. Therefore, these strategies like anchoring happened more in analytic evaluations which can be due to vague wording of scale. Regarding this, through analyzing TAPs data, he concluded that Translation Assessor interpreted some components

inconsistently because of some vagueness in the wording of the scale even though they showed similar patterns in assessment the translations.

Moreover, none of the Translation Assessor felt ambivalent in scoring holistically which indicated that Translation Assessor when scoring holistically are fundamentally in the center of their own cognitive process since they necessitate to come to a decision on one general score and not decide upon several separate scores. Therefore, it can be attributed to the conciseness of holistic assessment scale's components which help Translation Assessor in the decision-making process. Another strategy that one of the Translation Assessor adopted was reading the components and band descriptors reversely. Others frequently preferred to read the band scores of holistic assessment scale and components of analytic assessment scale in the sequence in which they appeared which agreed with the findings of Winke and Lim (2015).

The second finding of this study is that in their holistic assessments, it took Translation Assessor a long time to read, rate, and look for mistakes in the text because, as Barkaoui (2010a) explained, Translation Assessor needed time to rationalize their score. Therefore, they mostly gave their focus on their assessment not assessment scale' components. Only 1(R2) out of 6 Translation Assessor read the whole assessment scale (all band scores) and the rest only read one or went through the band scores till reached the appropriate one.

In analytic evaluations, the Translation Assessor rated the text faster but had difficulty making quick decisions when applying the components. Fast-decision making hinges on how much attention Translation Assessor paid to which components. It is worth noting that the Translation Assessor who adhered to the holistic scale components emphasized that reading components had an efficient role in their assessment and provided them with insight to avoid articulating biased score. Additionally, they perceived that having a concise synopsis of assessment criteria caused them to put their mind in consolidation which confirms the Terry' (1989) finding indicating that in the holistic scale, much attention is given to the message of the text not individual parts such as accuracy, mechanics, and complexity.

Diversity in the Translation Assessor' interactions with the holistic scale showed that each rater had his/her own assessment strategy. The Translation Assessor considered different aspects of translation such as grammar, organization, vocabulary, cohesion, and content proceeding the rhetoric and their general impression of the text. Further, according to their TAPs, after referring to the scale and giving evidence, it appeared that they paid more attention to organization and content components which is consistent with Shi (2001) who indicated that non-native translation Assessor generally emphasized more on content and organization whereas native Translation Assessor attended more to the language use when holistically assessment and scoring.

Moreover, variations in assessment processes are frequently attributed to individual differences in Translation Assessor' attentional focus (Cumming, 1990; Eckes, 2008). It is likely that the Translation Assessor focus on different assessment criteria in the translation when scoring or taking external and superficial features into account such as hand translation, indentation, and translation length which are not explained in the scale (Barkaoui, 2010a). In this study, the Translation Assessor put their attentional focus on all aspects of translation when assessment the translation holistically but did not consider most of them in scoring process in order to justify the assigned score and only judged based on substantial features like organization and content. This is in good agreement with Huot (1990) reports of a study on the assessment process in which the experienced Translation Assessor applied a holistic assessment scale. The findings showed that the Translation Assessor were highly influenced by content and organization. Findings from Plakans and Gebril (2017) and Winke and Lim (2015) also show that the disregard of organization component in the test taker' translation enhances the scoring variability and lessen the reliability of scoring of ESL/EFL translation. As a

result, when assessment translation using the holistic scale, Translation Assessor consider organization as an important component in EFL context.

Additionally, in the analytic scoring, Translation Assessor were in agreement on two components: organization and language use. This finding supported the results of Barkaoui (2010a) who asserted that Translation Assessor attached more importance to organization when they applied analytic scale to rate the EFL translations. Besides, the findings lend support to Winke and Lim (2015) who reported that Translation Assessor put high attention on organization component which led to the highest consistency. Li and He (2015) also mentioned that the organization received the most attention while assessment EFL translations analytically.

To add more, it was also found that in the analytic assessment scale, the Translation Assessor adhered to the scale's components even when they wanted to give evidence whereas in the holistic assessment they mixed their own criteria with the content of the scale or at least focused on two components: organization and content. Noticeably, this finding was also in line with those found by Barkaoui (2007) and Li and He (2015) who reported that the application of the analytic scale resulted in less frequent use of self- made criteria. In addition, in comparison with the holistic scale, Translation Assessor have a tendency to confine their attention to just those on the scale while assessment analytically (Goulden, 1994). In fact, the order of the components appearing in the analytic assessment scale (Jacobs, et al., 1981) does not reflect the importance of the components. In a study, Winke and Lim (2015) claimed that because content and organization are arranged first in the scale, Translation Assessor put more attention to these components. In this study, it was found that the Translation Assessor focused slightly less on the mechanics component in the analytic assessment; therefore, two Translation Assessor consulted the translation to re-rate the component while reading that component's criteria. During the translation assessment process, they might not examine mechanics errors or might not bear them in mind. This lack of consideration might stem from primacy effect (Underwood, 1975). The primacy effect brings a tendency to those Translation Assessor to ignore the mechanics component because this is listed last in the analytic scale and the Translation Assessor may well perceive it as the least important dimension. This finding is in line with the Winke and Lim (2015) who imply that, mechanics component received less attention among all analytic components.

Another finding of this study is that the Translation Assessor who considered the holistic scale components mostly focused on the strengths of the text while in analytic assessment they emphasized on the weaknesses which are considered to be of highly importance for decisions concerning promotion. This confirms the findings of Cumming (1990) and Elbow (1999) who claimed that assessment holistically frequently emphasizes what the writer does well.

In addition to the overall differences of perceptions regarding the use of assessment scales' components, the elements that are important in Translation Assessor' attention seem to vary a great deal as well. Experience assists Translation Assessor evaluate the text more easily because they internalize the assessment criteria as they gain experience (Barkaoui, 2010, 2011; Lumley, 2005). In this study, the element of experience cannot be considered as a reason for the application of scales components because some of the experienced Translation Assessor preferred not to utilize the scales particularly those who adopted their own idiosyncratic components. Translation Assessor were all highly experienced in the marking of translation assessments, but their behaviors towards the scales were totally different. As an example, R1 preferred to use holistic scale rather than analytic scale due to its time-saving advantage. This rater did not refer to the holistic scale components but one time at the end of his assessment process, after yielding score, in order to validate the assigned score. Even though this reference led him to revise the score, he preferred to use the scale and took advantage of its components. Therefore, assessment scale had a larger effect on the assessment process than assessment experiences (Barkaoui, 2010b).

Regarding the second research question, the Wilcoxon Signed Ranks Test showed that there was a difference between the Translation Assessor' scores assigned holistically and analytically. Studies in second/foreign language translation performance assessment showed that some factors influence the Translation Assessor' application of assessment scales. For instance, the scoring rubrics, Translation Assessor' educational-linguistic backgrounds, and teaching and assessment experience (Khodi, 2021; Huang, 2012; Li & Lindsey, 2015). Therefore, the quality of assessments is a central consideration in the interpretation of outcomes (Hamp-Lyons, 2007) and the different interaction regarding all these aspects of assessment can lead to different outcomes. Naturally, if one of these factors does not work or is not considered in the assessment process, there would not be a consistency among the scores. All Translation Assessor in this study were highly experienced in assessment and teaching translation. However, the wide range of different behaviors that the Translation Assessor showed concerning the scale components led to variations in the outcome.

Mumford and Attay (2021) characterized the factors that lead to discrepancy among Translation Assessor' scores: any rater types, assessment scales, interpretation of assessment scales, and the factors that make a text complicated to score. Different perceptions and interactions towards the scale' components can lead to the discrepancy as well. In other words, variations in assessment behaviors constantly were observed which affected the score that Translation Assessor assigned to the translation. Anchoring, ambivalence, disregard of components, and self-knowledge-based inference of components were some differences in Translation Assessor' perceptions.

In this study, it was also found that the Translation Assessor' outcomes in the analytic scoring were more lenient than holistic scores even though in the former, Translation Assessor fluctuated across five components and a wide range of sub-components. On the other hand, Deygers, Van Gorp, and Demeester (2018) found individually distinctive differences in rater severity among Translation Assessor in both analytic and holistic scorings. Adversely, the findings of the present study match well with Holzknecht et. al. (2018) findings who reported significant disagreement on the analytic and holistic assessment scores which implied that the Translation Assessor in their holistic assessments were more severe. Hunter, Jones, and Randhawa (1996) also indicated that in comparison with the holistic assessment, analytic scoring can result in slightly higher scores. Eckes (2008) claimed that inconsistency among equally-trained Translation Assessor is usually derived from their individual differences in their way of reading translation and considering the scale components they perceive as important. As a result, differences in the Translation Assessor' scoring behaviors might bring about low inter-rater agreement.

However, Lumley (2002) explained that different scoring behaviors do not fundamentally result in inconsistency among the Translation Assessor. In this study, although the Translation Assessor were equally prepared to apply assessment scales components, an inter-rater agreement was not observed among them which can further support previous findings in Choi (2002) that training does not necessarily affect the inter-rater reliability.

5. Conclusion

In sum, the findings of the present study showed that the Translation Assessor are predominately obsessed with content and organization components when using the holistic scale. In other words, in holistic assessment, the Translation Assessor mostly focused on the content and organization components while in the analytic assessment, the Translation Assessor considered all aspects of text such as length, lay-out, and capitalization. In addition, in holistic assessment, the Translation Assessor emphasized on the strengths of the translation whereas in the analytic scoring, the Translation Assessor put all their attention on the weaknesses of the text as well. Moreover, the findings revealed

that assessment scales components assist Translation Assessor to take into account all aspects of translation during the scoring process.

Overall, the findings showed that EFL translation Assessor will possibly be affected to various and unpredictable extents by scales components, translation criteria, and different interpretations that might have applied before to rate translations and frequently through their long experience. Thus, by making these components, interpretations, and criteria more examinable, definite, specific, and more valid results will be obtained.

The role of a scale is mostly to help challenge the diverse set of interactions Translation Assessor have between the text and the scale components while they rate the text; the interpretations they have might not be in accord with the orientation of the scale which may be derived from lack of explicitness of scales' components. Applying the behaviors identified in this study in Translation Assessor' training sessions can help them inform their own beliefs and priorities regarding the different types of assessment scales components and act effectively as well in assessment the texts.

Knowing the particular way that the Translation Assessor perceive the assessment task can improve the quality of assessment and lead to increasing the assessment accuracy and reliability of the scoring. The detailed interactions of the rater can also help identify the strategies Translation Assessor adopt in their assessments. In this study, some problems in the assessment process emerged which were due to the complexity and, vagueness of criteria in the analytic components which in turn led to adopting translation, comparison, and rereading of criteria in each component. In order to improve scoring consistency and efficiency and promote assessment transparency, the complexities related to the concepts and descriptors of the components should be solved and considered in training programs. In this way, the rater can do the assessment task in a consistent way.

Another implication of the study concerns the values that assessment scale developers attributed to different components of a scale. The results of the qualitative data on the scales in the assessment process showed how the Translation Assessor weighed the scales components such as content, organization, and mechanics. In fact, getting access to such list of priorities will assist training program organizers to be aware of how to develop the programs in a way that inform the Translation Assessor to cover and better understand equal value of all assessment components of analytic and holistic scales in the best possible way.

This study also suffers from a number of limitations. This study was carried out with a small sample size (i.e. nine Translation Assessor and one translation) in the particular EFL context of Iran. However, investigating these research questions with larger samples and in different contexts might create different findings. According to Barkaoui (2010a), TAPs do not reflect and show a whole picture of assessment process and affect both assessment process and outcomes as some Translation Assessor mentioned in their demographic information. Therefore, these intrinsic limitations of TAPs can affect the validity of the findings. According to Shohamy, Gordon, and Kraemer (1992), training plays an influential part in affecting Translation Assessor' behaviors using assessment scales to rate the text, particularly through clarification of assessment criteria. Due to the limitations caused by COVID-19 pandemic, many Translation Assessor were reluctant to cooperate in face-to-face sessions. Therefore, think-aloud sessions were conducted in a remote way which could affect the validity of the TAPs obtained.

Further studies can be conducted to identify the Translation Assessor' perceptions regarding the structurally similar scales. Addressing the process of assessment holistically and analytically in two time intervals is also required in order to shift the focus from the identification of the Translation Assessor' perceptions to the comparison of Translation Assessor' perceptions regarding the scales components. This study considered the assessment of only one sample translation. Future investigations can survey the identification of Translation Assessor' different perceptions on more translations. According to Charters (2003), a combination of interviews and TAPs would provide a more detailed picture of the participants' thought processes. Based on Qi (1998), a follow-up

interview might permit the participants to validate the interpretation or clarification of think-aloud sessions.

Acknowledgements

The current research was self-funding research. Therefore, the researchers would like to thank the Growingscholar Publisher who has reviewed and published this research.

References

Bacha, N. (2001). Translation evaluation: what can analytic versus holistic translation scoring tell us? System, 29, 371–383.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL translation scores and rater performance. Assessment in Education Principles Policy and Practice, 18(3):279-293.

Barkaoui, K. (2010a). Think-aloud protocols in research on translation assessment: An empirical study of their veridicality and reactivity. Language Testing, 28(1), 51–75.

Barkaoui, K. (2010b). Variability in ESL translation assessment processes: The role of the assessment scale and rater experience. Language Assessment Quarterly, 7(1), 54–74.

Barkaoui, K. (2007). Assessment scale impact on EFL translation marking: a mix-method study. Assessing translation, 12(2), 86-107.

Brown, A. V. (2009). Students' and teachers' perceptions of effective teaching in the foreign language classroom: A comparison of ideals and assessments. The Modern Language Journal, 93(1),46-60. https://doi.org/10.1111/j.1540-4781.2009.00827.x

Choi, Y. (2002). FACETS analysis of effects of rater training on secondary school English teachers' scoring of English translation. Journal of the Applied Linguistics Association of Korea, 18(1), 257-292.

Charters, E. (2003), The use of think-aloud methods in qualitative research an introduction to Think-aloud Methods. Brock Education, 12(2),68-82.

Coombe, C. (2010). Assessing foreign/second language translation ability. Education, Business and Society: Contemporary Middle Eastern Issues ,3(3),178-187.

Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry & research design: Choosing among five approaches. Thousand Oaks, California: SAGE.

Cumming, A., Kantor, R., & Powers, E. D. (2001). Scoring TOEFL translations and TOEFL 2000 protocol tasks: An investigation into Translation Assessor' decision making and development of a preliminary analytic framework. (TOEFL Monograph Series, Report No. 22.)

Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision making while assessment ESL/EFL translation tasks: A descriptive framework. Modern Language Journal, 86(21), 67–96.

Davidson, M., Howell, W. & Hoekerna , P.(2000). Effects of ethnicity and violent content on rubric scores in translation samples. Journal of Educational Research, 93(6).18-31.

Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 level and the dream of a common standard. Language Assessment Quarterly, 15(1), 44–58.

Eckes, T. (2006). Translation Assessor' perceptions of scoring criteria in translation and speaking performance assessments. Paper presented at the 28th Language Testing Research Colloquium (LTRC), Melbourne, Australia.

Eckes, T. (2008). Rater types in translation performance assessments: A classification approach to rater variability. Language Testing, 25(2), 155–185.

Elbow, P., (1999). Ranking, evaluating, and liking: sorting out three forms of judgments. In: Straub, R. (Ed.), A Sourcebook for Responding to Student Translation. Hampton Press, Inc, New Jersey, 175–196.

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing translation rater behavior on an academic English reading/translation test used for two purposes. Assessing Translation, 30, 21-31.

Hamp-Lyons, L. (1991a). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), Assessing second language translation in academic contexts. Norwood, NJ: Ablex. (pp. 241–276).

Hamp-Lyons, L. (2007).The Impact of Testing Practices on Teaching: Ideologies and alternatives. Cummins, J., Davison, C.(Ed). International Handbook of English Language Teaching - (pp.487-504). Springer.

Hijikata-Someya, Y.,Ono, M., Yamanishi,H.(2015).Evaluation by native and non-native English teacher Translation Assessor of Japanese students' summaries. English Language Teaching, 8 (7), pp. 1-12.

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived assessment scales for a story retelling speaking test. Language Assessment Quarterly, 10(4), 398–422.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of largescale ESL translation assessment. Assessing Translation, 17, 123–139.

Huot, B. (1990b). Reliability, validity, and holistic scoring: what we know and what we need to know. College Translation and Communication, 41(2), 201–213.

Hunter, D., Jones, R. M., & Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of translation. The Canadian Journal of Program Evaluation, 11(2), 61–85.

Kim,S, & Lee,H. K.(2015). Exploring Rater Behaviors During a Translation Assessment Discussion English Teaching, 70(1),97-121.

Khodi, A. (2021). The affectability of translation assessment scores: A G-theory analysis of rater, task, and scoring method contribution. Testing in Asia ,11,1-27. https://doi.org/10.1186/s40468-021-00134-5.

Knoch, U. (2007). The development and validation of an empirically-developed assessment scale for academic translation. Unpublished PhD, University of Aucklan.

Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hart®el, V.F. and Hughey, J.B. (1981). Testing ESL translation: a practical approach. Rowley, MA: Newbury House.

Joe, J., Harmes, J., & Hickerson, C. (2011). Using verbal report to explore rater perceptual processes in scoring: a mixed method application to oral communication assessment. Assessment in Education: Principles, Policy & Practice, 18, 239-259.

Jeong, H. (2019), Translation scale effects on Translation Assessor: an exploratory study, Jeong Language Testing in Asia, 9(20),1-19. https://doi.org/10.1186/s40468-019-0097-4

Li, H., & He, L. (2015) A Comparison of EFL Translation Assessor' Translation-Assessment Processes Across Two Types of Assessment Scales, Language Assessment Quarterly, 12(2), 178-212, DOI:10.1080/15434303.2015.1011738

Li, J. & Lindsey, P. (2015).Understanding variations between student and teacher application of rubrics. Assessing translation.26, 67-79.

Liu, Y. & Huang, J.(2020).The quality assurance of a national English translation assessment: Policy implications for quality improvement. Studies in Educational Evaluation,67(2), 1- 8. 10.1016/j.stueduc.2020.100941

Lumley, T. (2002). Assessment criteria in a large-scale translation test: what do they really mean to the Translation Assessor? Language Testing,19 (3) 246–276.

Lumley, T. (2005). Assessing second language translation: The rater's perspective. New York: Peter Lang.

McNamara, T. (1996). Measuring second language performance: Applied linguistics and language studies, ed. C. Candlin. London: Longman.
Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. Teaching and Teacher Education, 65, 48–60.
Mumford, S.& Attay, D. (2021). Teachers' perspectives on the causes of rater discrepancy in an English for Academic Purposes context. Assessing Translation,48,1-11.
Ohta, R., Plakans, L., & Gebril, A. (2018). Integrated translation scores based on holistic and multi-trait scales: A generalizability analysis. Assessing Translation,38, 21–36.
Ono, M., Yamanishi, H. Hijikata, & (2019). Holistic and Analytic Assessments of the TOEFL iBT® Integrated Translation Task. JLTA Journal,22 65–88.
Panadero, E. & Jonsson, A., (2020). A critical review of the arguments against the use of rubrics. Education Reseasrch Review, 20. 1-19. https://doi.org/10.1016/j.edurev.2020.100329
Plakans L. &Gebril (2017). Exploring the relationship of organization and connection with scores in integrated translation assessment. Assessing Translation,31, 98-112.
Qi, D. S. (1998). An inquiry into language-switching in second language composing processes. Canadian Modern Language Review, 54(3), 413-435.
Rahayu, E.Y. (2020). The Anonymous Teachers' Factors of Assessing Paragraph Translation. Journal of English for Academic and Specific Purposes.3(1),1-19.
Sakyi, A. A. (2003). A study of the holistic scoring behaviors of experienced and novice ESL instructors, University of Toronto, Toronto, Canada.
Sakyi, A. A. (2000). Validation of holistic scoring for ESL translation assessment: How Translation Assessor evaluate translation. In M. Milanovic, & A. J. Kunnan (Eds.), Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium (pp. 129–152). Cambridge: Cambridge University Press.
Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English translation. Language Testing, 18 (3) 303–32.
Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of Translation Assessor' background and training on the reliability of direct translation tests. Modern Language Journal, 76(1), 27–33. https://doi.org/10.2307/329895
Speck, B, & Jones, T. (1998). Direction in the grading of translation? What the literature on the grading of translation does and doesn't tell us. In F. Zak & C. C. Weaver (Eds.), The theory and practice of grading translation: Problems and possibilities (pp. 17-29). Albany, NY: SUNY Press.
Terry R.M. (1989). Teaching and evaluating translation as a communicative skill. Foreign Language Annals,22(1),43-52.
Turner, C. E., & Upshur, J. A. (2002). Assessment scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. TESOL.
Underwood, G. (1975). Perceptual distinctiveness and proactive interference in the primacy effect. Quarterly Journal of Experimental Psychology, 27(2), 289–294.
Wang,J., Engelhard, G., Raczynski, K., Song, T., Wolfe, E.D.(2017). Evaluating rater accuracy and perception for integrated translation assessments using a mixed-methods approach. Assessing Translation, 33, 36–47.
Weigle, S.C. (2002): Assessing translation. Cambridge, UK: Cambridge University Press. xiv, 268.
Weir, C. J. (2005). Language testing and validation: an evidence-based approach. Research and Practice in Applied Linguistics, 9(1), xiv, 301 https://doi.org/10.1057/9780230514577
Winke P., Lim, H. (2015), ESL translation cognitive Translation Assessor' processes in applying the Jacobs et. al. assessment scale: An eye movement study. Assessing Translation, 25, 38-54.