Journal of
Language and Translation
Volume 12, Number 3, 2022 (pp. 21-34)

Assessing Quality of Pedagogical Translations: Dominant Evaluative Methods in the Final Tests of Undergraduate Translation Courses

Hossein Heidari Tabrizi[1]*
Associate Professor, English Department, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran

Received: September 10, 2021 Accepted: November 8, 2021

ABSTRACT
Translation evaluation is an activity of vital concern in the context of translator education. However, it is one that is, despite being a common practice, under-researched and under-discussed. The leading purposes were to determine the dominant methods for translation evaluation of undergraduate students in the final tests of translation courses at Iranian universities and to identify its major characteristics. To meet the objectives, in a 30-minute semi-structured interview, 10 experienced university translation instructors responded to 10 questions on how they develop their translation tests, how they guarantee the quality of their tests, and most importantly, how they evaluate and score their students' performance. The results confirmed that the dominant method which is commonly and currently practiced is the purely essay-type format except at Payam-e-Noor University where multiple-choice items are always present as well. The findings also showed how much discrepancy exists among the Iranian translation teachers (as developers of final translation tests), who are least informed about the current translation evaluation methods. It was also revealed that the criteria they use for developing such tests and scoring student translations are not theory-driven but are highly subjective, mainly based on their personal experience and intuition. Hence, the quality and accountability of such tests are under serious question.

Keywords: Academic translation evaluation; Translation assessment; Translation evaluative practices; Translation quality; Undergraduate translation program

INTRODUCTION

Translation evaluation is an activity of vital concern and significance in the context of translator education; however, as stated by many scholars in the field around the turn of the century (Arango-Keeth & Koby, 2003; Bowker, 2000), it is one which is, despite being a common practice, as mentioned by Hatim and Mason (1997), "under-researched and under-discussed" (p. 197). In the debates on the subject of the assessment of translations in a Round-table discussion on translation in the New Millennium, McAlester (2003) even goes further: "This is an area in which Translation Studies has its worst failure" (p. 45). In sum, as elaborated by Hatim and Mason (1997), it can safely be concluded that in comparison "little is published on the ubiquitous activity of [translation] testing and evaluation" (p. 197). More recently, Tsagari and Van Deemter (2013, p. 11) "The field of measurement and assessment in translation and interpreting is growing but there is still much work to be done". At the beginning of the 2020s, the situation is almost the same (Heidari Tabrizi, 2021; Sun, Guzmán, & Specia, 2020; Yazdani, Heidari Tabrizi, & Chalak, 2020).

* Corresponding Author's Email:
heidaritabrizi@gmail.com

The reason for such neglect of this field of inquiry may be the fact that translation evaluation, though an extremely important issue in translation and translator training programs, is at the same time one of the most problematic areas of translation, having been referred to as a "great stumbling block" (Bassnett, 2013, p. 20), a "thorny issue" (Darwish, 2010, p. 99) a "most wretched question" (Malmkjaer, 1998, p. 70), a "chausse trappe" (Larose, 1998, p. xix), in the related literature. Translation evaluation schemes are also regarded as “dead ducks” (McAlester, 2003, p. 46) or “unsystematic, hit-and-miss methods” (Hatim & Mason, 1997, p. 198).

The principal difficulty surrounding translation evaluation as a tricky matter is its subjective nature: the notion of quality has such fuzzy and shifting boundaries, difficult to determine, that a translation which is deemed acceptable in one context or by one evaluator may be deemed inappropriate in another context or by other evaluators.

In other words, there is a consensus among scholars in the field of translation studies that neither is there a universally acceptable model of translation evaluation nor can the same set of objective criteria be applied uniformly to all translation activities (Bassnett, 2013; Drugan, 2013; Heidari Tabrizi, 2021; Honig, 1998a, 1998b; House, 2013; Tsagari & Van Deemter, 2013; Yazdani et al., 2020).

In academic setting, evaluating translations is even much more daunting because a translation teacher has an obligation to help students improve their performance. Teaching translation involves judging the quality of the translations produced by students and giving a score for the attainment of the intended objectives; that is, the instructional goals. As a matter of fact, translation evaluation through quality assessment is a fundamental part of the career of every translation teacher. There are all the time mid-term and final tests as well as other more formative diagnostic assessments done for pedagogical purposes in the academic institutes. Needless to say, every teacher of translation has an academic obligation to rank his/her students' work. In fact, translation teachers are said to play two major and simultaneous functions: they are both facilitators of learning and evaluators of what has been learnt. Thus, as stated by Honig (1998b) in educating translators, judging the translation quality "should not be an end but a means" (p. 32). On the whole, translation evaluation is undoubtedly one of the most difficult tasks facing a translation teacher: the problem of evaluation and decision-making in translation. It is unlikely that there will ever be a ready-made formula that will transform this task into a simple one; however, attempts have been made to investigate this issue from different perspectives (Drugan, 2013; Hatim & Mason, 1990, 1997; Honig, 1998a, 1998b; House, 1997, 2001a, 2001b; Munday, 2012; Sainz, 1994; Schaffner, 1998; Schiaffino & Zearo, 2005; Tsagari & Van Deemter, 2013; Waddington, 2001; Williams, 2004). As far as the present researcher knows such endeavors are exceptional in the academic Iranian environment.

In sum, it could possibly be claimed that the dominant trend for evaluating translation quality in academic settings in Iran is far behind the modern ones practiced in accredited universities throughout the world. One piece of evidence can be the frequent negative feedbacks teachers are likely to receive from the students about the final tests of translation in every semester. Still another piece of supporting evidence is the countless anecdotes one hears in professional conferences about the deficiencies of translation tests. Again, Honig (1998b) argues that “The least homogenous TQA criteria are assembled in university training course. The students feel that TQA is subjective and arbitrary, they try to adapt to the standards of teachers and they acquire neither self-awareness nor self-confidence” (p. 29).

In fact, the existing method commonly and currently used in the undergraduate translation program at Iranian universities does not seem to create the sense of satisfaction neither in the students nor in the teacher assessors themselves. This cardinal problem is exacerbated when such students sit for the MA Translation Entrance Tests where evaluation of students' competence in translation must be made in a systematic and highly valid as well as reliable yet practical way (Amiri Shalforoosh & Heidari Tabrizi, 2018; Heidari Tabrizi, Chalak, & Taherioun, 2014; Heidari Tabrizi & Pezeshki, 2015; Jalalpour & Heidari Tabrizi, 2017; Karimi, Heidari Tabrizi, & Chalak, 2016; Khalouzadeh, Heidari Tabrizi, & Chalak, 2013; Moeini Fard, Heidari Tabrizi, & Chalak, 2014; Montazer & Chalak, 2017; Shahsavarzadeh & Heidari Tabrizi, 2020; Valipoor, Heidari Tabrizi, & Chalak, 2019).

Having been teaching different courses of translation at a number of universities in Iran for more than 25 years, the researcher himself must confess that the reliability, validity and even in some cases the practicality of such tests as well as the way they are graded are under serious question. In most cases, translation students do not know on what criteria their work will be evaluated. Even much worse, some teachers and lecturers blithely substitute the authority of their position for any awareness of the complexity of the evaluative situations. The results are disastrous: students feel that the evaluation of their translations is done on the basis of arbitrary, subjective practices; they spend most of their energy adapting themselves to the personal non-objective criteria of their teachers and feel that it is a waste of time to gain insights into the nature of translation processes as provided by translation theories; consequently, they lack the self-awareness as well as the self-confidence they need to carry out translation tasks when they are on their own in the real — and confusing— world of translations.

Considering the relevant literature at hand as well the previous studies conducted by the researcher himself (Heidari Tabrizi, 2008, 2021; Heidari Tabrizi, Riazi, & Parhizgar, 2008; Yazdani et al., 2020), it can be claimed that translation teachers of Iranian universities are least informed and familiar, if at all, with the current translation evaluation approaches and methods in the field of translator education. This is in line with Honig (1998b, p. 29), "Obviously, many teachers and lecturers are not aware of the fact that there is such a wide variety of evaluation scenarios and applied criteria”. Likewise, Newmark (2003, p. 65) asserts that “… examination boards and examiners are not aware of the literature”.

The leading purposes of the present study were to determine the dominant trends/methods for translation evaluation of undergraduate students in final tests of translation courses at Iranian universities and to identify the major characteristics of such trends or methods encompassing test format, text choice, text difficulty, validity, reliability, scoring criteria, test rubrics, time allocation, and resources accessibility. In brief, the present study explored in depth the way translation teachers design, develop and prepare final tests as measures for checking on the quality of students’ academic translation in Iranian context. Accordingly, the present study aimed at finding appropriate answers to the following apparently simple, yet unresolved, central questions in assessing and evaluating translation: ‘what is to be assessed/evaluated’ and ‘how is it to be assessed/evaluated?’ These two general questions were then formulated and calibrated, in the context of academic translation at the Iranian universities, in the form of the following precise questions:

***Q1:** What are the dominant methods for translation evaluation of undergraduate students in final tests of translation courses at Iranian universities?*

***Q2:** What are the major characteristics of such trends/methods?*

METHOD

Design

The present study employed a qualitative design for categorizing and codifying the themes analytically explored through an in-depth interview with experienced translation teachers. The specification of the interviewees as well as those of the instrument itself are explained in detail in the following sections.

Participants

The participants included the translation teachers who attended the semi-structured interview sessions. To establish this sample, it should be mentioned that since the population of prospective interviewees were widely dispersed in different cities of Iran, the researcher, using the convenience sampling procedure, selected the sample from the following cities: Isfahan, Tehran, Shiraz, Bandar Abbas and Arak. First, the researcher checked the lists of full-time faculty members of English Departments of various universities offering the English Translation Program at the BA level. Then, those who were teaching translation courses there at the time of conducting the study were identified through consulting the Heads of the English Departments. Next, those who were more involved in teaching interpretation than in translation were excluded. Moreover, the teachers were required to have at least ten years of experience in teaching non-theoretical English translation courses at the university to be included in the final sample group for a semi-structured one-session interview. Finally, ten translation teachers qualified for the purposes of the present study accepted to participate in the interview. Table 1 summarizes the characteristics of the teachers who were finally interviewed by the researcher:

Table 1
Characteristics of the Interviewees

1- Sex	Male: 9	Female: 1	
2- Age	Range: 36-59	Average: 45	
3-Translation Teaching Experience		Range: 10-20	Average: 14
4-Affiliation	State: 3 IAU: 5	Payam-e-Noor: 1	Private: 1
5- Educational Background			

Major	Literature	TEFL	Tran.	Ling.	AZFA	Graduated from	
Level						Iran	abroad
B.A.	4	3	3			10	--
M.A.	1	6		3		10	
Ph.D.		4		1		3	2

Instrument

The instrument used in this study was the conventional one for typical qualitative research, that is, the semi-structured interview (Ary, Jacobs, Irvine, & Walker, 2019). The justification behind the application of this instrument but not others was that unlike other possible instruments like protocol analysis and portfolios, interview was more product-oriented in full harmony with the objectives of this research. Thus, to achieve the purposes of the present research study, the researcher developed and piloted a teacher-target interview in a systematic way to ensure the reliability as well as the validity of the instrument. Throughout the whole process of designing, preparing and conducting the interview as well as interpreting the outcome, the researcher carefully followed, among others, especially the guidelines and principles proposed by Ary et al. (2019) as well as the guidelines recommended by Cohen, Manion, and Morrison (2018) regarding the format, content, organization, sequencing, attractiveness, and comprehensibility of the instrument.

First of all, to guarantee the quality of the questions posed in the interview session, the researcher generated a shortlist of content specification through consulting the existing literature especially following the steps Bachman (1990) and McNamara (2000) proposed for test constructions. In brief, the questions posed mainly dealt with “why to test”, “what to test”, “how to test” and “when to test”: on how familiar they are with translation evaluation models currently in use, how they establish the test quality of their final translation tests, how they determine the time span to be allocated for such tests, how the papers are scored/evaluated and what consulting sources of information are allowed.

Moreover, as Klaudy (1996) who talks about “a human rights-based approach to correction of translations,” argues, “students have the right to know the evaluation system used to evaluate their translation, they have to know who is judging their work" (p. 200). This is also in line with Sainz (1994). Accordingly, one item was also allocated to this question: Would the testees be informed, through instructions, of how their translations are evaluated and scored? The questions were arranged in such a sequence that respondent's reactions to a question naturally led to posing the next one by the interviewer. One possible procedure for conducting the interview was to give participants the questions in advance a couple of days before the session to allow them prepare themselves for the interview. However, the researcher avoided such a procedure since the purpose of the study was to determine the status quo of the participants' knowledge, opinions, and attitudes about the translation evaluation and tests without preparation as such which would otherwise make the study biased.

As the final step, the questions were reviewed by three testing experts who unanimously approved their appropriateness and relevance. After applying the comments made by these experts for the improvement of the instrument, the researcher administered a trial interview session to pinpoint any possible problem with the practicality of the instrument. A copy of the finalized, refined version of the interview questions was approved once more by three other testing experts who unanimously ratified their fitness, relevance and consistency of the instrument to the purposes of the study.

To protect the privacy of participants while collecting, analyzing, and reporting data elicited from the interview, the researcher did his best to observe the ethical practice of confidentiality by separating any personal, identifying information provided by interviewees from the data. To mask their identity, every interviewee was assigned an identification number.

Data Collection Procedure

To collect the required data, ten male and female translation teachers having at least ten years of experience were invited to an in-depth 30-minute one-session semi-structured interview. The interview sessions were held by the researcher himself as the interviewer for each of the translation teachers separately. The allocated time for each session was about 30 minutes. The interviews were mainly conducted in English, but to ensure that the participants were able to express their ideas fully and clearly, they were allowed to use their native language (Persian) when necessary. The interview sessions were calibrated towards collecting rich, thick data on the interviewees' perceptions of translation evaluation in an academic setting, on how they develop their final tests of translation, establish their validity and reliability, and mark them. In so doing, a few predetermined, precise, clear and motivating questions were posed with considerable flexibility concerning follow-up questions pertinent to their teaching experiences.

In practice, by way of introduction, the participants were asked to complete a one-page questionnaire on their personal information: their sex, age (optional), educational background, teaching experience and the translation courses they had taught. Next, they were given the interview questions in writing to skim through for a couple of minutes. In this way, they got general information about what the interview was about which contributed to the structuredness of the interview. Then, in separate sessions for each participant, they were interviewed by the researcher himself. Of course, the interviews were done on different days in a time span of two months. The interview sessions were tape-recorded and then transcribed for further analyses. The interviewees were allowed to use their native language, that is, Persian in order to ensure that they can express themselves. However, just one of them preferred the interview to be conducted in Persian. It took at least 30 minutes and 45 minutes at most for each interview. All the interviews were tape-recorded to enhance the dependability of the data through techniques such as triangulation and member checks.

The interviewees were first asked about the degree of their familiarity with the existing modern quantitative and qualitative approaches, models and rating scales for translation scoring and evaluating in academic contexts. Then, they were asked to imagine they were going to develop a test for the final test of the course “Translation of Simple Texts.” The questions posed accordingly covered nine domains encompassing ‘Test Format’, ‘Text Choice’, ‘Difficulty Level’, ‘Validity’, ‘Reliability’, ‘Testee’s Awareness of Scoring Criteria’, ‘Instructions’, ‘Time Allocation’, and ‘Dictionaries/Glossaries’.

Data Analysis Procedure

The data analyses were done at two levels, namely data organization and data coding procedures utilized for the analysis of the data, as explained here. Before analyzing the data, it is absolutely essential, as a fundamental step in statistics and a key component of qualitative research, to organize the raw data into a manageable, easily understandable, more orderly form. In so doing, first, the recordings of the interview sessions were transformed into a textual form, that is, transcription. Then, the data collected through the transcription of the interviews were codified using the grid explained at the end of the previous section. To this end, the researcher and one of his colleagues with 20 years of experience in research data analysis got involved in identifying and codifying key topics in such data through studying them over a variety of times, looking for key ideas and labeling them by marginal notes and post-its. In fact, as the main categorizing strategy in qualitative research, the custom-made coding system created by the researcher for the purposes of this research study was applied to re-arrange the data into categories, domains, themes, or topics to facilitate the data comparison, data analysis, and drawing conclusions. In so doing, the researcher mainly followed the guidelines and principles proposed by Ary et al. (2019), Cohen et al. (2018), and Riazi (2016).

To contribute to the issue of reliability of the analyses made, the researcher did his best to establish the stability (intra-rater reliability) as well as the reproducibility (inter-rater reliability) of the coding scheme. In so doing, the interview transcriptions were analyzed and categorized by two coders independently using the same grid for coding. The coders were the researcher himself and one of his experienced colleagues well-accustomed to interviews thanks to more than twenty years of conducting language-related research. It is worth mentioning that the two coders operating independently were not working together to come to a consensus about what coding they would give.

Moreover, after a two-week interval, the same two coders re-coded the same data in the same way once more. Of course, to minimize any possible effects, if any at all, due to sequencing of the presentation of the tests, the researcher used the counterbalanced design: The sample tests were arranged in two opposite orders labeled ‘Form A’ and ‘Form B’. In the first phase of the coding process, ‘Form A’ was given to Coder One and ‘Form B’ to Coder Two while in the second phase the order was reversed. In cases where some inconsistency was observed between coders or between different codings of the same coder, the two coders discussed the case to reach a consensus. Accordingly, the inter-coder as well as intra-coder reliability coefficients were one; no need to use any statistical formulas.

RESULTS

To report and analyze the data in a systematic way, the findings and results reached as well as the analyses made are arranged and presented here in terms of the relevant data source and the sequence of the interview questions which were divided into two major sections. The first introductory part included just one general question:

General Introductory Question: *There are a variety of quantitative & qualitative approaches, models and rating scales for translation scoring & evaluating in academic contexts. Which ones are you familiar with?*

The answers given showed that except for just one single teacher, the majority of interviewees were not familiar with the existing approaches for translation evaluation at all, though they had been teaching translation courses for a long time. In other words, they applied their self-made criteria adopted based on their own experience and intuition rather than any kind of theory-driven sets of criteria. Even, two of the participants asserted that they do not believe in theories and theoretical frameworks as far as translation is concerned.

The second part of the interview covered more practical issues presented in nine questions. The translation teachers were asked to imagine they were going to develop a test for the final test of the course "Translation of Simple Texts". Then, nine questions were posed accordingly.

Interview Question One: *What kind of test format would you use? The MC items or Essay-type format or else?*

Eight participants responded that they preferred essay-type questions, that is, a passage of appropriate size over which there was no consensus among them; the size ranged from some passage(s) of about 200 words up to 750 or more. The other two preferred a combination of essay-type and multiple-choice items. One argued that multiple-choice items are appropriate to test the ability of students to choose best equivalents for a word or collocation. However, all agreed that since translation is a productive skill, multiple-choice items are not good for testing translation performance. They suggested that multiple-choice items can just measure the recognition, or at most the discrimination, ability which may be necessary in some translation tests but not sufficient.

Interview Question Two: *What kind of materials would you select? SEEN or UNSEEN texts?*

All the respondents believed that it is not an appropriate way to evaluate translation ability of students by giving them seen texts in the final achievement tests since they evaluate memory rather than translation. Students usually memorize the translation given by the teacher in the class and restore it in the test. Even one of the teachers argued that students usually get a lower grade on tests with seen texts since they lose their concentration on what they are doing. However, three teachers mentioned that they sometimes include as a source of warm-up encouragement and motivation for the students one or two seen texts in their tests as well.

Interview Question Three: *How would you adjust the difficulty level of the test?*

Since the texts included in such tests are mostly unseen; that is, new to students, it is of great importance to determine and adjust their difficulty level. Again, the interviewees mentioned that they followed no objective criteria or formula to do so. Rather, they are heavily dependent on their intuition and experience. They did not believe that any readability formula can be of any help in this regard. But the majority of the participants do believe that the vocabulary and the grammatical structures used in the text can indicate its difficulty level for the students. Six teachers also added that for each course they had a pack of texts usually written by the same writer on one single topic or a collection of different parts of a single passage. They would give their students some of these and put aside a couple or more for final tests or other tests. In this way, they tried to make sure that the texts used in the test(s) had a high correlation with those translated in the class as far as the difficulty level is concerned.

Interview Question Four: *How would you establish the validity of the test, especially its content validity? Some table of test specifications?*

The responses revealed that none of the participants used the table of test specifications strongly recommended by test developers to guarantee the test validity. They just relied on their intuition and experience for developing a valid test. Most of the participants argued that they select especially the unseen texts in such a way that the text be similar to those translated in the classes in terms of vocabulary level, grammatical patterns, subject matter and style. Thus, since the texts reflect those done in the classes, the validity of the final test, they claimed, would be established in this way.

Interview Question Five: *What about its reliability? In other words, what kind of criteria* *would you use for evaluating & scoring student translation?*

Seven participants mentioned that they use a “red-ink-scribble-over-the-TT” approach; that is, the penalty system and deduction of scores for errors. In other words, they were mainly concerned with the microstructure of the texts; namely, translation at word and sentence levels including students' choice of equivalents and appropriateness of the structures used. They mentioned that they assigned weighted score to translation problems or traps of the texts and grades for major and minor errors were deducted from a perfect score. In contrast, the three other argued for a more holistic way of evaluating and scoring the translation taking into consideration the macrostructures of the texts. However, none used a standard reliable scoring rubric; in fact, they were totally unaware of the existing rating scales.

While four interviewees asserted that they used their own translation as a necessary, though not sufficient, criterion for scoring the student translations in the final tests, the six others argued that they did not usually compared the student translations with their own. Of course, four out of these mentioned that instructor's own translations may be regarded as one of the possible solutions for the translation problem but not as the final translation.

Interview Question Six: *Would you inform the testees of how their translations are evaluated and scored?*

All the participants claimed that their students are mostly informed about how the teachers evaluate and score their translations either through the in-test instructions or by the prior in-class explanations given by the teachers during the course. They argued that for a valid reliable evaluation, it is necessary for the teachers to inform their students about the way their translations are considered.

Interview Question Seven: *Would you write instructions for the different parts of the test? If yes, how?*

At first, all the participants claimed that they did write instructions for their tests. But further investigation revealed that eight used repeatedly just a fixed cliché form of instructions: "Translate the following into proper Persian." One interviewee argued that it is nonsense to use such a qualifier as 'proper' since the students are supposed to provide an acceptable translation which has no way but to follow proper standards of Persian.

Interview Question Eight: *How would you allocate the appropriate amount of time needed for individual test tasks or for the entire test?*

All the interviewees indicated that they determine the amount of time required for a test again by their own personal experience and intuition rather than on a standard objective basis. Of course, they suggested some clues in so doing. All believed that the length of the passage(s) is a good criterion for allocating the time needed. Six proposed the time it takes the teacher himself or herself translate the text(s) as a reliable basis for estimating the time by giving the students an amount of time two or three times of that took by the teacher. One participant even asserted that he developed all his tests for a 90-minute session.

Interview Question Nine: *Would you allow the testees to use dictionaries and/or glossaries/terminologies? Why?*

Almost all of the participants except for one admitted that students should be allowed to consult information resources especially general dictionaries and technical glossaries and terminologies in the test sessions. The main justification they proposed was the fact that in daily real-life situations such references are normally and typically available to the translator. Moreover, translation tests are not intended to measure mere vocabulary knowledge of the testee at all. The only interviewee who was somehow against this view argued that instead of allowing such references it is better to provide the testee with the dictionary definition of the key difficult words at the bottom or next to the text(s) of course in the source language. He believed that such references when allowed may result in lack of concentration; the testee consumes most of the time allocated for the test on checking a wide number of words usually much more than enough.

DISCUSSION

In inquiries like this, it is always difficult to reach conclusions and find implications based on the findings since nothing has been manipulated or controlled by the researcher and therefore the safest thing would be to report or in fact describe simply the phenomena as they are. Nevertheless, even a piece of research of a highly qualitative nature has to come up with some tangible outcomes. In fact, the thick, rich data elicited from the teacher-target interview provided the researcher with enough data from different perspectives to be able to come to solid enough conclusions regarding the research questions based on the findings, analyses and results presented, categorized and analyzed here. To present the conclusions reached and the interpretations made, the researcher has tried to explain his own hopefully justifiable responses to each question and then assess them in relation to the existing body of the literature.

The findings elicited from the research instrument; that is, the teacher-targeted interviews, revealed that the dominant trend for translation evaluation of undergraduate students in translation courses at Iranian universities is mainly not formative but the summative evaluation conducted at the end of the semesters. In other words, most part of the student final score is allocated to the final test of the course. The main features of this method explored and found through this piece of research are further presented and discussed below.

As for the test format, it was revealed that in such tests, the dominant method commonly and currently practiced in the undergraduate translation program at Iranian universities is the purely essay-type format except at Payam-e-Noor University where multiple-choice items are always present as well. In practice, some texts are given to be translated (mostly to the testees' native language) with time limitation. Moreover, the evaluation is mainly concerned with just the product of the translation process rather than the process itself.

The results obtained from the data gathered by the teacher-target interview revealed that teacher-assessors select the texts for the tests on a rather subjective basis: they choose the test materials especially the 'unseen' texts by their intuition or experience. As for the 'unseen' materials, they usually select some text of similar topic for the test from the same source from which the texts for the class activities are selected. They argue that since the texts are produced by the same writer on similar topics, they are of similar difficulty level as well. In fact, no standard criteria are used by the teacher-assessors to determine objectively whether the texts are appropriate for the students and the objective of the course as far as the difficulty level, subject-matter and length of the texts are concerned. Translation experts believe that the length, the topic, the diction and the linguistic (structural) complexity of the texts can be useful in determining their difficulty level. They reject the adequacy of readability formulas in this regard, however.

Accordingly, the validity of these tests in general and their content validity in particular is under serious question. As the results of the teacher-target interview showed, the correspondence between what is taught in the class and what is tested in the final achievement tests of translation is subjectively checked by the teachers as test-developers based on what they think and feel. As for the content validity, the test content must be bound to the content of the instruction constrained, in turn, by the instructional objectives; the test must include a proper representative sample of the course material. However, the content relevance as well as the content coverage of the test materials is not objectively determined in the translation tests currently in use at the Iranian universities since the teachers do not follow a standard procedure for test specification or blue print in a systematic way. This is in line with the findings of Heidari Tabrizi (2021) and those of Yazdani et al. (2020).

The results of the interview sessions of the teachers also revealed that the scoring methods currently used by English translation teachers at Iranian universities are mostly based on the so-called 'Classical True Score Measurement

Theory'. Thus, no rating scale is at work in practice. In other words, the rater consistency as well as the task consistency is not checked at all. The scoring is done on a subjective basis usually through holistic, impressionistic method. Hence, it can safely be concluded that the translation teachers follow no certain standard models in scoring the translation of their students. Accordingly, the researcher should side with researchers such as Honig (1998b), who argues that most teachers are not aware of the wide variety of models and criteria applied for translation evaluation throughout the world. Similarly, Newmark (2003) criticizes the test-takers for being unaware of the literature on translation evaluation schemes. All in all, the reliability of these tests is questionable too.

As for the test instructions and directions, the findings indicated that the Iranian translation teachers, like experts in the field of language testing, believe that the testees must be informed of how to perform the translation tasks and how their performance is to be evaluated and scored. They recommended that to do so, explicit written test instructions be developed for the tests. Moreover, the teachers interviewed emphasized that they did write instructions for their tests to inform the testees. However, further scrutiny and analysis showed that while the tests did have instructions in most cases, unfortunately they were limited to the general statement 'translate the following into [proper] Persian'. This cliché is not enough at all. As Bachman (1990) suggests, test instructions, as one of the facets of test rubric, are of paramount importance in testees' performance. They must specify in themselves how testees are expected to proceed in taking the test. He argues that test instructions should inform testees on "the conditions under which the test will be taken, the procedures to be followed and the nature of the tasks they are to complete" (Bachman, 1990, p. 123). In fact, it is the instructions that undertake most of the responsibilities for setting the testees' expectations and appropriately motivating them to show their best performance on the test. The findings of Heidari Tabrizi et al. (2008) showing that the attitude of the translation students towards the present test directions also approves their inadequacy in helping the translation testee with how to perform on the final tests.

Time allocation, as one of the facets of the test rubric, is another characteristic of any testing method. The method used for translation evaluation of undergraduate students in final tests of translation courses at Iranian universities is no exception. Again, the responses the teachers provided for the relevant question in the interviews supplied enough supporting pieces of evidence to conclude that no systematic procedure or formulas are followed by the test-developers; no logical pattern is behind the allocation of the time required. It was recommended in the literature that the amount of time required should be allocated according to the length of the translation tasks and the time the teachers themselves spend to translate the test texts. However, the research results showed in most cases these suggestions are not taken into consideration at all.

The testees' access to information resources during the test must be explained as far as the dominant method for translation evaluation of undergraduate students in final tests of translation courses at Iranian universities is concerned. The pieces of evidence elicited by the teacher-target interview showed that teachers believe the testees should have access to dictionaries and glossaries during the tests. They add that in cases where the testees are not permitted to use these resources, they should be provided with definitions of some trouble-making words at the either sides or at the bottom of the test booklet. In addition, translation experts (such as Newmark, 1988) argue that to guarantee the authenticity of translation job and to avoid artificiality, in all translation tests students must be allowed to use dictionaries during their test since they can always consult human/non-human resources and references especially a dictionary during translating a text in their normal career as a translator.

Surprisingly enough, the curriculum of 'English Translation Program' approved by the Supreme Council of Programming of the then-

Ministry of Culture and Higher Education (now officially known as Ministry of Science, Research, and Technology) currently in use in Iran remains absolutely silent on the issue of evaluating the student translations in general and their translations in the final tests in particular. In other words, no pieces of information can be found in the syllabi developed in the curriculum for translation courses as for designing, preparing, administering and scoring the final tests required. Thus, the teachers had no choice but to rely on their very own standards, models and criteria.

CONCLUSION

As the findings of the present study showed, the Iranian undergraduate students majoring in English translation feel strongly dissatisfied with the majority of test-quality aspects of their final tests. Accordingly, it is strongly recommended that translation teachers when developing the final tests try to improve the validity, reliability and practicality. They should apply reasonable criteria in selecting texts of appropriate length, topic and difficulty level as well as in allocating the sufficient time for final tests. To guarantee the authenticity of the tests, it is also advisable that the students be allowed to consult general and technical information resources. Moreover, the test instructions should be written in a much more instructive way, providing the testees with vital information on what points they should observe in translating test texts and how their translations are scored. It is also suggestible to make more use of appropriate modern-day test forms and formats. Discrepancy exists among the Iranian translation teachers (as developers of final translation tests), who are least informed with the current translation evaluation methods. It was also revealed that the criteria they use for developing such tests and scoring student translations are not theory-driven but are highly subjective, mainly based on their personal experience and intuition. Hence, the quality and accountability of such tests are under serious question. Moreover, the overall impression of the final tests on the students proved to be rather negative.

By way of a conclusion to the present piece of research, as a touchstone, the researcher enumerates these general guidelines uncovered as follow: First of all, a shift must be made towards more direct, performance-based methods of testing and evaluation.

Accordingly, it is obvious that to develop a translation evaluation scheme and scoring rubrics is by no means an easy task due to the large variety of factors affecting its success or failure in actual use. The near-to-ideal rating model for evaluating translation performance in academic contexts must, more than anything else, consist of a large set of consistent well-defined criteria. As a part of solution to this problem, the developers of the rating scales should be selective; they have to find a way to limit the number of criteria to be used and the number of processes to be done into a manageable, practical proportion. Instead, however, the present researchers propose that the criteria to be used must be prioritized by the teacher-evaluator who applies the rating scale just using those which are most relevant to the situational, cultural context and the course requirement.

Thus, the potential translation evaluation scheme for the Iranian context may consist of a list of criteria composed of a manageable number of items selected based on the ideas proposed by translation scholars in the literature to date. Two types of criteria can be introduced into the evaluation scheme. On one hand, there should be micro-criteria which focus on more language-oriented microstructures of the translation task such as form, accuracy, mechanics of writing, grammatical points, lexical equivalence, style, shifts and error types. On the other hand, there should be macro-criteria which include more socio-pragmatic macrostructures such as function, fluency, naturalness, cohesion, coherence, genre and register.

As far as task development process is concerned, teachers must combine their logic, personal experience and intuition with the more systematic standard procedures found in the literature. According to Chalhoub-Deville (2001, pp. 214-217), a good task should have the following characteristics. As for the rater selection, characteristics and consistency, since almost always the translation teacher plays the role of the scorer as well, it is not possible to establish or check the inter-scorer reliability. In addition, scoring is done just once by the teacher-evaluator due to the numerous papers s/he has to score; hence, no intra-scorer reliability coefficient can be computed either. As such, the teachers must be highly familiar and trained to use the scoring rubrics either through pre-service education or in-service training; their overall expertise in this regard is of paramount importance. Last but not least, translation teachers must always keep in mind that, as Weber (1984) asserts, student translations in final tests should never ever be scored and evaluated in such a way as if they are ready-to-be-published piece of work.

REFERENCES

Amiri Shalforoosh , E., & Heidari Tabrizi, H. (2018). The study of English culture specific items in Persian translation based on House’s model: The case of Waiting *for Godot. International Journal of English Linguistics, 8*(1), 135-145.

Arango-Keeth, F., & Koby, G. S. (2003). Assessing assessment: Translator training evaluation and the needs of industry quality assessment. In B. J. Baer (Ed.), *Beyond the ivory tower. Rethinking translation pedagogy* (pp. 117-134). Amsterdam/Philadelphia: John Benjamins.

Ary, D., Jacobs, L., Irvine, C., & Walker, D. (2019). *Introduction to research in education 10th edth ed*. Boston (MA): Cengage Learning.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bassnett, S. (2013). *Translation Studies* (4th ed.). London: Routledge.

Bowker, L. (2000). A corpus-based approach to evaluating student translations. *The Translator, 6*(2), 183-210.

Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 210-228). Harlow, UK: Pearson Education.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). London: Routledge.

Darwish, A. (2010). *Translation applied!: An introduction to applied translation studies-A transactional model*. Melbourne: Writescope Publishers.

Drugan, J. (2013). *Quality in professional translation: Assessment and improvement*. London: Bloomsbury.

Hatim, B., & Mason, I. (1990). *Discourse and the Translator*. London: Routledge.

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

Heidari Tabrizi, H. (2008). *Towards developing a framework for the evaluation of Iranian undergraduate students’ academic translation.* Doctoral thesis, Shiraz University, Iran.

Heidari Tabrizi, H. (2021). Evaluative practices for assessing translation quality: A content analysis of Iranian undergraduate students' academic translations. *International Journal of Language Studies, 15*(3), 65-88.

Heidari Tabrizi, H., Chalak, A., & Taherioun, A. H. (2014). Assessing the Quality of Persian Translation of Orwell's Nineteen Eighty-four Based on House's Model: Overt-covert Translation Distinction. *Acta Linguistica Asiatica, 4*(3), 29-42.

Heidari Tabrizi, H., & Pezeshki, M. (2015). Strategies used in translation of scientific texts to cope with lexical gaps (Case of Biomass Gasification and Pyrolysis Book). *Theory and Practice in Language Studies, 5*(6), 1173-1178.

Heidari Tabrizi, H., Riazi, A., & Parhizgar, R. (2008). On the translation evaluation methods as practiced in Iranian universities' BA translation program:

the attitude of students. *TELL, 2*(7), 71-87.

Honig, H. G. (1998a). Complexity, contrastive linguistics and translator training: Comments on responses. In C. Schaffner (Ed.), *Translation and quality* (pp. 83-89). Clevedon: Multilingual Matters Limited.

Honig, H. G. (1998b). Positions, power and practice: Functionalist approaches and translation quality assessment. In C. Schaffner (Ed.), *Translation and quality* (pp. 6-34). Clevedon: Multilingual Matters Limited.

House, J. (1997). *Translation quality assessment: A model revisited*. Tubingen: Gunter Narr Verlag.

House, J. (2001a). How do we know when a translation is good? In E. Steiner & C. Yallop (Eds.), *Exploring translation and multilingual text production* (pp. 127-160). Berlin: De Gruyter Mouton.

House, J. (2001b). Translation quality assessment: Linguistic description versus social evaluation. *Meta*, *46*(2), 243-257.

House, J. (2013). How do we know when a translation is good? *Exploring translation and multilingual text production* (pp. 127-160): De Gruyter Mouton.

Jalalpour, E., & Heidari Tabrizi, H. (2017). A study of English translation of colloquial expressions in two translations of Jamalzadeh: once upon a time and Isfahan is half the world. *Journal of Language Teaching and Research*, *8*(5), 1011-1021.

Karimi, M., Heidari Tabrizi, Hossein , & Chalak, A. (2016). Challenges in English to Persian translation of contracts and agreements: The case of Iranian English translation students. *Journal of Applied Linguistics and Language Research, 3*(6), 188-198.

Khalouzadeh, E., Heidari Tabrizi, H., & Chalak, A. (2013). Translation of news texts in Persian political magazines: van Dijk’s model of critical discourse analysis. *Journal of Translation Studies*, *10*(40), 67-76.

Klaudy, K. (1996). Quality assessment in school vs. professional translation. In C. Dollerup & V. Appel (Eds.), *Teaching translation and interpreting 3: New horizons* (pp. 197-203). Amsterdam/ Philadelphia: John Benjamins

Larose, R. (1998). Méthodologie de l'évaluation des traductions. *Meta, 43*(2), 163-186.

Malmkjaer, K. (1998). Linguistics in functionland and through the front door: A response to Hans G. Honig. In C. Schaffner (Ed.), *Translation and quality* (pp. 70-74). Clevedon: Multilingual Matters Limited.

McAlester, G. (2003). Comments in the ‘Round-table discussion on translation in the New Millennium’. In M. R. G. M. Anderman (Ed.), *Translation today: Trends and perspectives* (pp. 13-51): Multilingual Matters Limited.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Moeini Fard, Z., Heidari Tabrizi, H., & Chalak, A. (2014). Translation Quality Assessment of English Equivalents of Persian Proper Nouns: A case of bilingual tourist signposts in Isfahan. *International journal of foreign language Teaching and Research, 2*(8), 25-34.

Montazer, E., & Chalak, A. (2017). Interpretation strategies used by Iranian tour guides in translating cultural specific items. *Journal of Applied Linguistics and Language Research, 4*(8), 121-132.

Munday, J. (2012). *Evaluation in translation: Critical points of translator decision-making*. London: Routledge.

Newmark, P. (1988). *A textbook of translation*. New York: Prentice hall New York.

Newmark, P. (2003). No global communication without translation. In G. M. Anderman & M. Rogers (Eds.), *Translation today: Trends and* *perspectives* (pp. 55-67). Clevedon: Multilingual Matters Limited.

Riazi, A. M. (2016). *The Routledge encyclopedia of research methods in applied linguistics*. New York: Routledge.

Sainz, M. J. (1994). Student-centred corrections of translations. In C. Dollerup & A. Lindegaard (Eds.), *Teaching Translation and Interpreting 2* (pp. 133–141). Amsterdam/Philadelphia: John Benjamins.

Schaffner, C. (Ed.) (1998). *Translation and quality*. London: Routledge.

Schiaffino, R., & Zearo, F. (2005). *Translation quality measurement in practice.* Paper presented at the Proceedings of the 46th Annual Conference of the American Translation Association.

Shahsavarzadeh, S., & Heidari Tabrizi, H. (2020). Investigating translation theories course in Iranian universities: Students' expectations and perceptions in focus. *Research in English Language Pedagogy, 8*(1), 167-194.

Sun, S., Guzmán, F., & Specia, L. (2020). *Are we Estimating or Guesstimating Translation Quality?* Paper presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Tsagari, D., & Van Deemter, R. (2013). *Assessment issues in language translation and interpreting*. Frankfurt am Main: Peter Lang AG.

Valipoor, K., Heidari Tabrizi, H., & Chalak, A. (2019). Cultural-specific items in translation of the Holy Quran by Irving. *Linguistic Research in the Holy Quran, 8*(1), 43-52.

Waddington, C. (2001). Different methods of evaluating student translations: The question of validity. *Meta, 46*(2), 311-325.

Weber, W. K. (1984). *Training Translators and Conference Interpreters. Language in Education: Theory and Practice*. New York: Harcourt Brace Jovanovich.

Williams, M. (2004). *Translation quality assessment: An argumentation-centred approach*. Ottawa: University of Ottawa Press.

Yazdani, S., Heidari Tabrizi, H., & Chalak, A. (2020). Exploratory-cumulative vs. disputational talk on cognitive dependency of translation studies: Intermediate level students in focus. *International journal of foreign language Teaching and Research, 8*(33), 39-57.

Biodata

Hossein Heidari Tabrizi is an associate professor of TEFL and Head of the Graduate Department of English at Islamic Azad University, Isfahan Branch, Isfahan, Iran. He is the director-in-charge of Research in English Language Pedagogy (RELP) published at IAU, Isfahan Branch and was selected as the top researcher of the English Department in 2016 and 2020. His research interests include Language Assessment, Translation Studies, and Critical Discourse Analysis.

Email: *heidaritabrizi@gmail.com*