Revista de Lingüística y Lenguas Aplicadas
Vol. 14 año 2019, 1-18
EISSN 1886-6298
https://doi.org/10.4995/rlyla.2019.11068

LOGISTIC CALIBRATED ITEMS (LCI) METHOD: DOES IT SOLVE SUBJECTIVITY IN TRANSLATION EVALUATION AND ASSESSMENT?

Alireza Akbari
University of Isfahan, Iran

Abstract: *This research paper introduces a model of translation evaluation called Logistic Calibrated Items method. The aim of LCI method is to maximize a translators' performance and to identify top competent translators through detecting all parsing items within a source text. Parsing items are extracted by the application of Brat software. The purpose of LCI was to identify parsing items having optimal item difficulty and item discrimination values. LCI method involves six stages: (1) holistic scoring; (2) the application of Brat software to extract all parsing items; (3) the calculation of item difficulty; (4) the calculation of item discrimination; (5) the identification of items with optimal item difficulty and item discrimination values; and (6) the recalculation of scores. 125 translation students and 4 professional translation evaluators took part in this research. The final results showed that LCI method was more consistent when compared to holistic method. Limitations and implications were also discussed.*

Keywords: *translation evaluation product, Logistic Calibrated Items method, item difficulty, item discrimination, recalculation of scores.*

1. INTRODUCTION

Traditionally, translation evaluation has been based on detecting and analyzing errors, a methodology which requires a remarkable investment in human resources when assessing a huge volume of translation drafts both in academic and professional settings (Schmitt, 2005). Thus far, research in the field of translation evaluation and assessment has predominately associated with descriptive and theoretical aspects and has concentrated on the following themes namely; criteria for good and poor translations (Newmark, 1991), the nature of translation and language errors (Gouadec, 1989), pragmatic and linguistic translation quality assessment (Nord, 2005), analyzing translation quality assessment based on text linguistic analysis (Larose, 1989), determining different textual levels and associating the significance of mistakes with these levels (Dancette, 1989), and various levels of translation competence (Stansfield *et al.,* 1992).

In the context of the above, the field of Translation Studies (hereafter TS) has vindicated the need for more experimental evidence for the assessment/evaluation of different translation tests (multiple-choice translation tests, expository translation exams, etc.) (Akbari and Segers, 2017b). Although educational and professional organizations have executed 'the certification of translation skills' (Eyckmans *et al.,* 2013) according to different test administrations, the validity (e.g. concurrent validity, statistical conclusion validity) and reliability of them remain questionable. Therefore, translation evaluation and assessment is by some means depended on the codes of practice rather than empirical explorations (ibid.). The field of translation evaluation covers the following themes such as translation process, translation product (the target text), translation service (e.g. client, complaints, and invoicing), and translation competence accordingly. Nonetheless, the mentioned themes cannot be evaluated/assessed/measured in the same way and necessitates different means of evaluation approaches.

According to Anckaert *et al.* (2008), there is no evaluation method which links the 'performance indicator' to the fundamental competence in a psychometric fashion. Lack of test development to assess translation competence

To cite this article: Akbari, A. (2019). "Logistic Calibrated Items (LCI) Method: does it solve subjectivity in translation evaluation and assessment?" *Revista de Lingüística y Lenguas Aplicadas*, *14*, 1-18. https://doi.org/10.4995/rlyla.2019.11068
Correspondence author: bahariakbar2020@gmail.com

BY-NC-ND
Crossref Similarity Check Powered by iThenticate

Received: 2018-11-29 Accepted: 2019-03-26

can be explained by two factors: (1) translation tests are not sufficiently valid and authentic to measure language ability (competence) for which this resulted in a definite overlooking of popularity during the era of Communicative Approach (CA) (Widdowson, 1978). This may be caused by the fact that translation tests are not laid open to 'the same psychometric scrutiny as other language testing formats' (e.g. cloze-test) (Eyckmans *et al.,* 2013); (2) There exists an epistemological aperture between the human science (e.g. translation and interpreting studies (T/I studies), language and literature, language and linguistics, etc.) and hard science (biology, chemistry, etc.). The surmise that it is impossible to objectify the quality of translation while covering its very essence may be very resolute among language instructors and translation scholars/trainers/instructors whose 'corporate culture exhibits a marked reticence towards the use of statistics' (Eyckmans *et al.,* 2012, 2013). In this direction, testing translation and interpreting approaches/skills have been more or less carried out by practitioners rather than of translation scholars and researchers. Thanks to different psychometric approaches, a fair amount of research in the field of reliability and validity of language tests has been recognized. Nonetheless, the field of T/I studies has been fallen back and requires more scrutinies.

Generally, translation evaluation focuses on issues namely; (i) translation process, (ii) translation, (iii) translator, (iv) linguistic, textual, and paralinguistic competences (Kockaert and Segers, 2017), and situatedness (Muñoz Martín, 2010). According to educational and professional settings, translation evaluation can be performed as maintained by a criterion-referenced approach (Schmitt, 2005) and can be gauged in respect of some 'assessment grids' to make translation evaluation more valid and reliable. These contexts can never sufficiently lessen the degree of subjectivity in translation evaluation. What's more, the scoring system which is susceptible to be influenced by contrast effect menaces the reliability and validity of translation tests. Contrast effect refers to 'a magnification or diminishment of perception as a result of previous exposure to something of lesser or greater quality, but of the same base characteristics' (González, 2018).

In the light of the above, the present research is an attempt to introduce an objective translation evaluation method called Logistic Calibration Items (LCI hereafter). LCI method is specified by the total number of parsing items within a source text in accordance with translation relevance and norm-criterion referenced assessment methods. The term 'calibration' was used to characterize docimologically justified parsing items. Docimology alludes to the theory of the art of testing and docimologically justified items refer to such items which have optimal corrected item-total correlations (r_{it} value hereafter) and appropriate item difficulty (p-value) based on 1-Parameter Logistic Model (1-PL) of item response theory (IRT) (the main target of LCI method). As is the case with the Preselected Items Evaluation (PIE) (Kockaert and Segers, 2017), correct and incorrect solutions are categorized for each parsing item in the source text in LCI method. LCI method consists of the following stages: (i) holistic scoring system based on evaluators' impression/intuition/anticipation; (ii) the use of Brat Visualization software Stanford CoreNLP parser to detect each parsing item within a source text; (iii) the calculation of item difficulty by 1-Parameter Logistic Model; (iv) the calculation of r_{it} value; (v) the extraction of docimologically justifed and calibrated items which have optimal r_{it} ($r_{it} \geq 0.30$) and item difficulty (positive reliability coefficient and p-value < 0.05) values; and consequently (vi) the re-calculation of scores.

2. STATE OF THE ART

2.1. A Review of Translation Evaluation

Translation evaluation is predominately represented by a criterion-referenced assessment. In terms of educational and professional settings, assessment and evaluative grids are used in an effort to make translation evaluation more reliable and valid. Conventionally, evaluative grids encompass 'a near exhaustive taxonomy of different kinds of mistakes or bonuses' (e.g. word choice, stylistic conventions, text coherence and cohesion) (Eyckmans *et al.,* 2013). Although the use of the evaluative grids is caused by an evaluator's desire to take into account different dimensions of translation competence, one has to affirm that these dimensions are unable to lessen the degree of subjectivity in translation evaluation and assessment (Anckaert *et al.,* 2008). Apart from the subjective features of translation evaluation sub-competences, other elements may menace the reliability and validity of translation administration tests. Let us start with an evaluator's consistency throughout the task of translation scoring during a specific time interval. Not only will the system of scoring be susceptible to a contrast effect, it is also compulsory to put forward a "sound testing practice" which discriminates good testing items from the bad ones. Moreover, all obtained scores must be docimologically acceptable and the scoring system must differentiate the standard quality of translations. For that reason, scholars and researchers from the field of translation quality assessment and evaluation (Conde Ruano, 2005; Kockaert and Segers, 2017; Akbari and Segers, 2017a) are now accrediting issues such as the degree of inter-intra-rater reliability (rater variability), construct, concurrent, and statistical conclusion validity as well as ecological validity. The methodology of educational and professional measurements along with the standpoint of theories of language testing are being transferred to the field of translation evaluation product and translation quality assessment in order to reach reliable and valid methods/approaches to measure translation competence. With that in mind, the dominant purpose of the present research paper is to emancipate translation evaluation from extraneous and irrelevant variables which have an effect on the outcome of assessment (Eyckmans *et al.,* 2013).

BY-NC-ND

2.2. Different Translation Evaluation Models

2.2.1. Holistic Method of Translation Evaluation

Holistic method of evaluation is supposed to be an objective and accurate method of translation evaluation (Kussmaul, 1995). This method of evaluation has a short range of objectivity and resiliency due to an evaluator/corrector's anticipation/impression and the type of translation errors that students make throughout their translations. Actually, this method as 'attending to the text as a whole' (Hamp-Lyons, 1991:246) has been employed very differently by translation instructors and evaluators. Holistic method of translation evaluates the overall quality of a translation in terms of a translator's impression (Mariana *et al.,* 2015). This method is considered fast yet fully *subjective* since it is depended on the taste of an evaluator while scoring a translation. As Kockaert and Segers (2017:149) contend, 'the value judgments of different holistic evaluators on the same translation can vary greatly'. For example, one evaluator takes into account one translation as excellent and acceptable; while another grader considers the same translation as fair or unacceptable (Eyckmans *et al.,* 2012). Garant (2001) has put forward that 'point-based error focused grading' as a type of paradigm shift has been superseded by holistic method of assessment at the University of Helsinki. According to Kockaert and Segers (2017), translation is suitably evaluated based on 'discourse level holistic evaluation' rather than 'grammar-like and analytical' evaluations. This method also concentrates mainly on a 'context-sensitive evaluation' (Akbari and Segers, 2017b) and is surmised to shift from 'exclusive attention to grammatical errors in translation tests' (Kockaert and Segers, 2017:149).

Although holistic method of evaluation supposes to determine higher inter-coder (inter-rater) reliability (Barkaoui, 2011) and lead to 'produce reliable and consistent assessment' (Cumming *et al.,* 2002:67), this method is 'not necessarily an indicator of the raters actually applying the scale in a consistent way' (Harsch and Martin, 2013). Higher degree of inter-rater reliability will conceal differences among evaluators 'for the criterion scores' (ibid.) and menace the degree of validity. According to Weigle (2002:114), 'holistic scoring has also come under criticism in recent years for its focus on achieving high inter-rater reliability at the expense of validity'. In line with Weigle, Barkaoui (2010:516) has pointed out that evaluators can move away from 'the criteria originally designed to define what is being assessed'. Therefore, 'this can reduce score consistency across and within raters and, ultimately, change the meaning of the scores' (ibid.).

Even though this method is acceptable, evaluators/graders are not always in a position of agreement when scoring translations. This fully shows that this method does not have sufficient objectivity. According to Bahameed (2016:144), holistic method of evaluation depends to a certain degree on 'the corrector's personal anticipation and appreciation'. By the same token, there do not exist any specific criteria available while scoring a translation based on holistic method. Bahameed (2016) continues that this method can never simply specify the top students as their scores 'may reach one-third out of the whole translation class'. This is due to the fact that students are not responsible for minor mistakes (e.g. lexical, grammatical, etc.). These minor mistakes can never be disregarded by a grader since they initiate a matter in the quality of holistic evaluation that is too arduous to measure. The leniency of holistic evaluation can negatively reverberate on the quality of a translation and eventually on the teaching process (Akbari and Gholamzadeh Bazarbash, 2017). In the context of the above, holistic method of translation evaluation is not supportable.

2.2.2. Analytic Method of Translation Evaluation

Analytic method of translation evaluation (also it is called evaluation grids method) is associated with error analysis and is maintained to be more valid and reliable in comparison with holistic method (Waddington, 2001:136). In analytic method of evaluation, an evaluator maintains a grid which includes a number of error levels and types. Therefore, the number of error types and error levels can be increased; nonetheless, this must be done with prudence. The reason is that an augment in error levels and error types can minimize the feasibility of this method. Analytic method evaluates the overall quality of a translation by investigating the text segments (e.g. individual words, clauses, paragraphs, etc.) in accordance with a certain number of criteria such as omission, addition, misinterpretation, and so forth. Eyckmans *et al.* (2013) have pointed out that translation errors should be identified with regard to 'the evaluation grid criteria'. Furthermore, an evaluator should firstly ascertain the types of errors such as errors associated with translation or language and s/he must maintain the pertinent information in the margin with regard to the nature of errors accordingly.

To put it in a nutshell, a translator has 'a better understanding of what is right and what is wrong in translation' (Kockaert and Segers, 2017:150) in analytic method. Analytic method has a disadvantage that an evaluator focusing on the small text segment of the source text does not definitely have an exhaustive prospect of the target text. This method is also *subjective* and necessitates more time compared to holistic method.

2.2.3. Preselected Items Evaluation (PIE) Method

The Preselected Items Evaluation (PIE) method is a system of scoring which was devised by Hendrik. J. Kockaert and Winibert Segers. This method is suitable for summative assessment. The purpose of a summative assessment is 'to evaluate student learning at the end of an instructional unit by comparing it against some standards or

BY-NC-ND

benchmarks' (Eberly Center, 2016). In terms of functionality and time management, PIE method only investigates the number of preselected items within a source text. PIE method is considered to be a calibrated method since it checks the accuracy of 'the measuring instrument'. Additionally, this method is called a dichotomous method because it vets the discrimination between correct and incorrect solutions (Kockaert and Segers, 2017:150). This calibrated and dichotomous method which does not distinguish between levels of errors, is suitable for all language combinations and comprises the following stages: (i) the preselection of a limited number of items in a source text based on evaluators' expertise; (ii) the identification of correct and incorrect solutions to the preselected items; (iii) the calculation of p-docimology [cf. p-value] or item difficulty (the proportion of examinees provides a correct solution for an item); (iv) the calculation of d-index or item discrimination (candidates' differentiations on the basis of the items being measured) based on the method of extreme group; and (v) the recalculation of scores in terms of optimal p-docimology and d-index values (PIE run scores). According to Lei and Wu (2007), the calculation of p-value and d-index associates with 'the minimum number of items needed for a desired level of score reliability or measurement accuracy'. The ideal range of p-value in PIE method 'should be higher than 0.20 and lower than 0.90' (Kockaert and Segers, 2017). With this in mind, the larger the sample size of the participants provides a correct solution for an item, the easier the selected item will be. To calculate item discrimination (d-index), PIE method employs the method of extreme group ('analysis of continuous variables sometimes proceeds by selecting individuals on the basis of extreme scores of a sample distribution and submitting only those extreme scores for further analysis') (Preacher *et al.,* 2005) through the application of 27% rule. That is to say, this method identifies the top 27% of candidates and the bottom 27% of candidates of the entire score ranking. The application of 27% rule will maximize differences in a normal distribution (Wiersma and Jurs, 1990). Items are preselected in terms of translation brief relevance and test specific criteria. Having administered a test, an evaluator calculates the difficulty of the selected items according to p-value and d-index. Items which are not responding to docimological standards (poor p-value and d-index) will be removed from the test.

Several studies have been conducted about PIE method to fine-tune it and proves its capacity to *objectify* the evaluation of translation products (Kockaert and Segers, 2014, Kockaert and Segers, 2017, Akbari and Segers, 2017a,c,d). These articles mainly discuss PIE method and its application for English, Dutch, French, and Persian languages and analyze its reliability (no sign of validity) (Akbari and Segers, 2017b) compared to holistic and analytic evaluations. Nonetheless, more research must be carried out to identify and solve specific scientific aspects. For instance, the validity of PIE method has not so far been recorded and has been under critical questions. No vindications are proposed of why items within a source text are preselected as the most difficult or simple items. On what basis does an evaluator preselect an item within a text? What are the appropriate number of preselected items in a source text? Once a translation is evaluated, what happens to other mistakes in a text? And consequently, what typology of items must be preselected within a source text (linguistic items, extralinguistic items, etc.) when applying PIE method?

3. METHODOLOGY

3.1. Description of the Participants and Study Conditions

This research paper took place in 2018. 125 translation students from the Bachelor of Arts in Translation Studies at the University of Sheikhbahaei and the University of Isfahan took part in this research. The participants were all Persian native speakers (L1) with average age of 22 years. They all passed the obligatory courses related to the literary, political, journalistic, translation of legal documents, and medical translation through which they were subjected to different translational texts. The participants were requested to translate a short text from English (L2) to Persian (L1). Even though the participants varied in their level of English language proficiency, the standard preconception was that it was by and large of good standard, since the registration in their study programs necessitated proof of passing compulsory credits such as medical, legal, economic, and political translation courses. The participants were asked to translate a short political text into Persian (L1) (see Appendix 1). They were all acquainted with political styles, terminologies, and structures as they passed the necessary courses related to the political translation. The length, type, and the difficulty of the selected political source text were considered indicative for the materials taught in the translation courses at the University of Isfahan and Sheikhbahaei University. Eventually, four representative translations made by four official translation agencies were provided for the graders; therefore, they would have access to different yet correct equivalents once evaluating and scoring translations. These official translation agencies had long-established experience (approximately 8 years) in assessing and translating political texts and documents.

3.2. Procedures

The provided translations were handed to 4 translation evaluators and were asked to commonly score the drafts. The evaluators were selected in terms of their longstanding experience (nearly 10 years) in translation quality assessment and evaluation. They were selected from the University of Isfahan and Sheikhbahei University. At first sight, there were asked to score the translations holistically based on the Waddington's framework (2001) (see Appendix 2). The reason to ask the evaluators to score holistically was just to show the difference between holistic

BY-NC-ND

scoring and LCI scoring at the end of the research. The scores made by holistic and LCI scoring systems were all calculated up to 20. Having scored translations holistically, the evaluators were asked to identify parsing items with optimal item difficulty by the application of 1-Pramater Logistic Model of IRT (positive reliability coefficient and p-value < 0.05) (Stata, 2016) and then to detect discriminating items based on optimal r_{it} values (0.30 and above for good and very good items) (SPSS, 2018). Moreover, the evaluators were apprised about the quasi-experimental design of the present research.

4. THE ADMINISTRATION OF LCI METHOD: FROM HOLISTIC CALCULATION TO SCORE RECALCULATION

- Stages of LCI Method

(I) Holistic Scoring of Translations

As mentioned, 125 translation students from two well-known universities were asked to translate a short political text (L2) into simple Persian (L1). When the translation task was done, four professional translation evaluators were assigned to commonly score the translation drafts based on the Waddington's (2001) framework. The participants' scores were as follows:

Table 1. Holistic Scoring (Docimologically Unjustified Scores).

Par	Score (Holistic)/20	Par	Score (Holistic)/20	Par	Score (Holistic)/20	Par	Score (Holistic)/20	Par	Score (Holistic)/20
1	13	26	10	51	13	76	15	101	16
2	15	27	12	52	15	77	15	102	15
3	15	28	13	53	14	78	17	103	17
4	16	29	13	54	10	79	15	104	14
5	14	30	17	55	12	80	16	105	18
6	10	31	10	56	15	81	13	106	19
7	11	32	20	57	16	82	19	107	14
8	13	33	20	58	17	83	12	108	15
9	12	34	18	59	17	84	15	109	15
10	17	35	18	60	17	85	16	110	15
11	15	36	9	61	12	86	17	111	18
12	20	37	11	62	12	87	15	112	15
13	18	38	13	63	10	88	14	113	12
14	14	39	14	64	15	89	13	114	10
15	12	40	14	65	15	90	15	115	16
16	11	41	15	66	15	91	14	116	13
17	12	42	12	67	18	92	15	117	14
18	12	43	13	68	18	93	14	118	19
19	11	44	12	69	20	94	17	119	13
20	15	45	11	70	20	95	12	120	15
21	16	46	15	71	18	96	13	121	20
22	16	47	15	72	17	97	14	122	20
23	17	48	15	73	18	98	15	123	12
24	15	49	15	74	18	99	14	124	16
25	10	50	13	75	18	100	14	125	15

According to Table 1, participants [6], [25], [26], [31], [54], [63], and [114] obtained the lowest scores compared to the rest of translation participants. According to the evaluators' comments and remarks, those participants largely applied literal translation (word-for-word translation), which end in ambiguous target text meanings for some parts of the source text. In this direction, those participants did not adapt optimal approaches when translating the source text. Moreover, they in question commit major semantic errors which caused their translations to a great extent ambiguous, unclear, and inaccurate. Those participants on many occasions lost the contextual function of the source text and resorted to word-for-word translation.

BY-NC-ND

(II) The Application of Brat Visualization Stanford CoreNLP Parser

Brat CoreNLP is a web-based tool which annotates and makes texts into parses. This tool was designed for structured annotation based on different NLP (Natural Language Processing) tasks. The aim of the Brat tool is to 'support manual curation efforts and increase annotator productivity using NLP technique' (Stenetorp *et al.,* 2012).

Modern annotation and parsing tools are technically-directed and many of them present 'little support to users beyond the minimum required functionality' (ibid.). In this respect, tools with user-friendly interfaces can support human decisions and help to provide the quality of annotations or parsers as well as making them more accessible to non-technical users. Furthermore, these tools ameliorate parsing productivity and functionality. As a parsing tool, Brat software is based on previously STAV text annotation visualizer. STAV text annotation visualizer was devised to assist users to acquire an exhaustive comprehension of convoluted annotations including a great number of various 'sematic types, dense, partially overlapping text annotations, and non-projective sets of connections between annotations' (Stenetorp *et al.,* 2011). This tool is thoroughly configurable and can be established to authenticate most text parsing tasks. Furthermore, the Brat software has been used to generate well-over 50,000 parses in thousands of documents. The purpose of LCI method for using the Brat Visualization CoreNLP software was to determine every parse or annotation (norm-referenced evaluation) within the source text and consequently identify such parses which were docimologically justified (criterion-referenced evaluation) (parses with optimal item difficulty and item discrimination). This tool categorizes every parse into specific classifications such as JJ (adjectives), NNS (common noun plural form), CC (coordinating conjunctions), NNP (proper nouns), MD (modal verb), VBN (verb past participles), IN (prepositions), amod (adjectival modifier), nmod (noun modifier), nsubjpass (passive nominal subject), aux (auxiliary), auxpass (auxiliary passive), and so forth. Figure 1 shows an extraction of the source text annotated by the Brat CoreNLP. Having imported the whole source text, the Brat software automatically exported 257 annotations (basic dependencies) based on the neighboring parses within a source text. At this phase, the four professional evaluators tried to extract all the annotations in a source text and then compared them to the participants' translations. This comparison was conducted as per the representative translations.

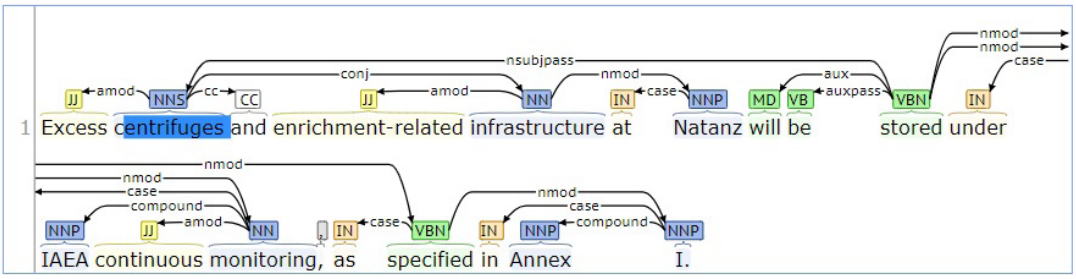

Figure 1. Brat Annotation Tool.

(III) The Calculation of Item Difficulty (1-Parameter Logistic Model)

The statistical models which are applied in the Item Response Theory (IRT) suppose that a participant's probability of answering an item correctly is associated with the participant's ability [cf. competence] and the features of an item being measured. The participant's competence is considered the main feature of the person which is called competence parameter. This competence parameter is believed to be considered a fundamental and unobservable latent trait which assists an individual to provide a correct answer for an item. The application of 1-Parameter Logistic Model (1-PL) (also known as item difficulty or threshold parameter) is to determine and measure the degree of difficulty of items. 1-PL measures and evaluates the location of an item within a continuum. As Umobong (2017:129) puts forward,

> The item parameter is believed to be a continuum with the upper end indicating greater proficiency in whatever is measured than the lower end. This means that items located towards the right side of the continuum demands an individual to possess greater proficiency (ability) in order to answer correctly, than items located towards the left side of the continuum.

This research paper used Stata ver.16 to calculate the difficulty of each parsing item. Under 1-PL, the probability of correct items that they differed regarding their degree of difficulty were determined. Due to space limitation, this paper only brought acceptable items which had positive reliability coefficient and p-value < 0.05 (95% confidence interval) (Stata, 2016). Mind that, such items in this section are not considered an absolute license for score recalculation. Therefore, the next stage is to calculate item discrimination. In this light, items which have optimal range of p-values (item difficulty) and appropriate amount of r_{it} values are considered docimologically justifiable items and must be included for the recalculation of scores. The purpose of 1-PL is to find out which parsing item is considered a simple item and which one is considered a difficult item.

BY-NC-ND

Table 2. Acceptable Degrees of Items' Difficulty Based on 1-PL (α = 0.05).

	Reliability Coefficient	Std. Err.	Z	P < 0.05
ITEM 9	0.6135796	0.20727	2.96	0.003
ITEM 11	0.4636934	0.2036581	2.28	0.023
ITEM 12	0.5757467	0.2062605	2.79	0.005
ITEM 15	1.053717	0.2238635	4.71	0.000
ITEM 17	0.4636934	0.2036581	2.28	0.023
ITEM 20	0.5381685	0.2053231	2.62	0.009
ITEM 23	0.5381685	0.2053231	2.62	0.009
ITEM 24	0.500824	0.2044561	2.45	0.014
ITEM 30	0.7288296	0.2107522	3.46	0.001
ITEM 32	0.500824	0.2044561	2.45	0.014
ITEM 38	0.4636934	0.2036581	2.28	0.023
ITEM 50	0.1001953	0.1992226	0.50	0.005
ITEM 52	0.9282814	0.2182241	4.25	0.000
ITEM 54	1.011341	0.2218769	4.56	0.000
ITEM 56	0.9695425	0.2199988	4.41	0.000
ITEM 57	0.9282814	0.2182241	4.25	0.000
ITEM 59	0.8472268	0.2149666	3.94	0.000
ITEM 60	0.6900976	0.2095137	3.29	0.001
ITEM 61	0.9282814	0.2182241	4.25	0.000
ITEM 74	0.6900976	0.2095137	3.29	0.001
ITEM 79	0.500824	0.2044561	2.45	0.014
ITEM 81	0.7679104	0.2120719	3.62	0.000
ITEM 83	0.6135796	0.20727	2.96	0.003
ITEM 84	0.8472268	0.2149666	3.94	0.000
ITEM 89	0.5757467	0.2062605	2.79	0.005
ITEM 90	0.6900976	0.2095137	3.29	0.001
ITEM 91	0.6135796	0.20727	2.96	0.003
ITEM 93	0.5381685	0.2053231	2.62	0.009
ITEM 94	0.6135796	0.20727	2.96	0.003
ITEM 95	0.500824	0.2044561	2.45	0.014
ITEM 97	0.8472268	0.2149666	3.94	0.000
ITEM 98	0.6516889	0.2083537	3.13	0.002
ITEM 99	0.4636934	0.2036581	2.28	0.023
ITEM 103	0.7288296	0.2107522	3.46	0.001
ITEM 105	0.4636934	0.2036581	2.28	0.023
ITEM 106	0.8073666	0.2134757	3.78	0.000
ITEM 108	0.4267574	0.2029275	2.10	0.035
ITEM 109	0.5757467	0.2062605	2.79	0.005
ITEM 110	0.4636934	0.2036581	2.28	0.023
ITEM 111	0.500824	0.2044561	2.45	0.014
ITEM 120	0.6135796	0.20727	2.96	0.003
ITEM 122	0.4636934	0.2036581	2.28	0.023
ITEM 123	0.5757467	0.2062605	2.79	0.005
ITEM 126	1.053717	0.2238635	4.71	0.000
ITEM 128	0.4636934	0.2036581	2.28	0.023
ITEM 131	0.5381685	0.2053231	2.62	0.009
ITEM 134	0.5381685	0.2053231	2.62	0.009
ITEM 135	0.500824	0.2044561	2.45	0.014
ITEM 141	0.7288296	0.2107522	3.46	0.001
ITEM 143	0.500824	0.2044561	2.45	0.014
ITEM 149	0.4636934	0.2036581	2.28	0.023

Table 2, continues on the next page

BY-NC-ND

Table 2, continues from the previous page

	Reliability Coefficient	Std. Err.	Z	P < 0.05
ITEM 161	0.6900976	0.2095137	3.29	0.001
ITEM 163	0.9282814	0.2182241	4.25	0.000
ITEM 165	1.011341	0.2218769	4.56	0.000
ITEM 167	0.9695425	0.2199988	4.41	0.000
ITEM 168	0.9282814	0.2182241	4.25	0.000
ITEM 170	0.8472268	0.2149666	3.94	0.000
ITEM 171	0.6900976	0.2095137	3.29	0.001
ITEM 172	0.9282814	0.2182241	4.25	0.000
ITEM 185	0.6900976	0.2095137	3.29	0.001
ITEM 190	0.500824	0.2044561	2.45	0.014
ITEM 192	0.7679104	0.2120719	3.62	0.000
ITEM 194	0.6135796	0.20727	2.96	0.003
ITEM 195	0.8472268	0.2149666	3.94	0.000
ITEM 200	0.5757467	0.2062605	2.79	0.005
ITEM 201	0.6900976	0.2095137	3.29	0.001
ITEM 202	0.6135796	0.20727	2.96	0.003
ITEM 204	0.5381685	0.2053231	2.62	0.009
ITEM 205	0.6135796	0.20727	2.96	0.003
ITEM 206	0.500824	0.2044561	2.45	0.014
ITEM 208	0.8472268	0.2149666	3.94	0.000
ITEM 209	0.6516889	0.2083537	3.13	0.002
ITEM 210	0.4636934	0.2036581	2.28	0.023
ITEM 214	0.7288296	0.2107522	3.46	0.001
ITEM 216	0.4636934	0.2036581	2.28	0.023
ITEM 217	0.8073666	0.2134757	3.78	0.000
ITEM 219	0.4267574	0.2029275	2.10	0.035
ITEM 220	0.5757467	0.2062605	2.79	0.005
ITEM 221	0.4636934	0.2036581	2.28	0.023
ITEM 222	0.500824	0.2044561	2.45	0.014
ITEM 231	0.6135796	0.20727	2.96	0.003
ITEM 233	0.4636934	0.2036581	2.28	0.023
ITEM 234	0.5757467	0.2062605	2.79	0.005
ITEM 237	1.053717	0.2238635	4.71	0.000
ITEM 239	0.4636934	0.2036581	2.28	0.023
ITEM 242	0.5381685	0.2053231	2.62	0.009
ITEM 245	0.5381685	0.2053231	2.62	0.009
ITEM 246	0.500824	0.2044561	2.45	0.014
ITEM 252	0.7288296	0.2107522	3.46	0.001
ITEM 254	0.500824	0.2044561	2.45	0.014

In this light, item 176 was considered the least difficult item compared to the whole items with regard to the reliability coefficient (-1.129195) and the p-value (0.000 < 0.05). As noted earlier, an item must have a positive reliability coefficient and the value lower than 5% to be considered an optimal item difficulty (Stata, 2016). Although item 176 has a value lower than 0.05, it has a negative reliability coefficient and it cannot be included as an optimal indicator for item difficulty. On the other hand, item 15 was taken into account the most difficult item in terms of the reliability coefficient (1.053717) and the p-value (0.000 < 0.05).

(IV) The Calculation of R_{it} Value (Item Discrimination)

R_{it} value (also known as corrected item-total correlation as well as item discrimination) is used to reverberate 'the performance of the item versus the test as a whole' (van Antwerpen, 2016). This value informs a researcher/scholar to what degree an item assists to single out good participants (higher scorers) and weak participants (lower scorers) from the entire pool of test takers. Simply put, the application of r_{it} value shows the discriminating properties of an item. Moreover, r_{it} value tells a researcher that to what extent items are correctly answered by high-performing participants compared to low-performing participants (positive discrimination index [between

BY-NC-ND

0 and 1]). On the other hand, if the majority of low-performing participants choose a correct answer for an item compared to high-performing participants, then the item being measured has a negative discrimination index (between -1 and 0). By the same token, Eyckmans and Anckaert (2017) have noted that:

> The r_{it} value calculates the correlation between the item and the rest of the scale, without that item being considered as part of the scale; that is, it reflects the amount the item contributes to the test's global reliability.

In order to detect the discriminating items, a researcher has to identify items with good discriminating power ($r_{it} \geq 0.30$). In this vein, items with a r_{it} value of 0.40 and above are indicators of very good items; items with a r_{it} value of 0.30 to 0.39 are considered good discriminators; items with a r_{it} value of 0.20 to 0.29 have fairly good discriminatory power; and consequently items with a r_{it} value of 0.19 or less are considered poor discriminators (Anckaert *et al.,* 2008). As a matter of fact, only items with good and very good discriminatory power are included within a test (Kockaert and Segers, 2017; Akbari and Segers, 2017a; Eyckmans and Anckaert, 2017). This research identified all items with fair and poor discriminatory power and excluded them from the test since their range of r_{it} values were inferior to 0.30 (Table 3).

Table 3: Excluded Items from the Test ($R_{it} < 0.30$).

R_{it} Value (Corrected Item-Total Correlation)					
Items	*Scale Variance if Item Deleted*	*Corrected Item-Total Correlation*	*Items*	*Scale Variance if Item Deleted*	*Corrected Item-Total Correlation*
ITEM 6	3332.265	0.291	ITEM 137	3333.525	0.282
ITEM 7	3334.639	0.255	ITEM 151	3334.550	0.251
ITEM 8	3334.313	0.256	ITEM 153	3333.061	0.277
ITEM 40	3334.550	0.251	ITEM 154	3332.509	0.286
ITEM 42	3333.061	0.277	ITEM 155	3348.903	0.003
ITEM 43	3332.509	0.286	ITEM 156	3349.794	-0.012
ITEM 44	3348.903	0.003	ITEM 157	3338.844	0.177
ITEM 45	3349.794	-0.012	ITEM 159	3343.018	0.105
ITEM 46	3338.844	0.177	ITEM 160	3344.519	0.079
ITEM 48	3343.018	0.105	ITEM 161	3337.813	0.204
ITEM 49	3344.519	0.079	ITEM 164	3347.957	0.020
ITEM 50	3337.813	0.204	ITEM 169	3335.960	0.227
ITEM 53	3347.957	0.020	ITEM 189	3334.811	0.247
ITEM 58	3335.960	0.227	ITEM 196	3335.455	0.236
ITEM 78	3334.811	0.247	ITEM 199	3332.361	0.289
ITEM 85	3335.455	0.236	ITEM 213	3350.794	-0.029
ITEM 88	3332.361	0.289	ITEM 215	3333.271	0.275
ITEM 102	3350.794	-0.029	ITEM 218	3333.351	0.274
ITEM 104	3333.271	0.275	ITEM 228	3332.265	0.291
ITEM 107	3333.351	0.274	ITEM 229	3334.639	0.255
ITEM 117	3332.265	0.291	ITEM 230	3334.313	0.256
ITEM 118	3334.639	0.255	ITEM 248	3333.525	0.282
ITEM 119	3334.313	0.256			

(V) The Identification of Docimologically Justifed and Calibrated Items

Based on the obtained results from 1-PL (item difficulty) and r_{it} values (item discrimination) in stages III and IV, 88 parsing items were categorized as docimologically calibrated parsing items. From the ninety acceptable items in 1-PL, 2 parsing items (parsing items 50 and 161) were not docimologically calibrated due lower r_{it} values (below 0.30); therefore, they were excluded from the test. The rest had acceptable item difficulty and item discrimination. The complete list of the docimologically accepted parsing items is as follows:

BY-NC-ND

Table 4. Docimologically Calibrated Parsing Items (Criterion-Referenced Evaluation).

Item	*P-value*	R_{it}	Item	*P-value*	R_{it}	Item	*P-value*	R_{it}
9	0.003	0.543	38	0.023	0.391	81	0.000	0.541
11	0.023	0.718	52	0.000	0.316	83	0.003	0.610
12	0.005	0.419	54	0.000	0.375	84	0.000	0.664
15	0.000	0.647	56	0.000	0.344	89	0.005	0.703
17	0.023	0.607	57	0.000	0.342	90	0.001	0.614
20	0.009	0.584	59	0.000	0.470	91	0.003	0.644
23	0.009	0.612	60	0.001	0.474	93	0.009	0.691
24	0.014	0.612	61	0.000	0.477	94	0.003	0.433
30	0.001	0.461	74	0.001	0.468	95	0.014	0.554
32	0.014	0.495	79	0.014	0.394	97	0.000	0.616

Item	*P-value*	R_{it}	Item	*P-value*	R_{it}	Item	*P-value*	R_{it}
98	0.002	0.752	122	0.023	0.718	163	0.000	0.316
99	0.023	0.634	123	0.005	0.419	165	0.000	0.375
103	0.001	0.397	126	0.000	0.647	167	0.000	0.344
105	0.023	0.389	128	0.023	0.607	168	0.000	0.342
106	0.000	0.390	131	0.009	0.584	170	0.000	0.470
108	0.035	0.478	134	0.009	0.612	171	0.001	0.474
109	0.005	0.499	135	0.014	0.612	172	0.000	0.477
110	0.023	0.500	141	0.001	0.461	185	0.001	0.468
111	0.014	0.402	143	0.014	0.495	190	0.014	0.394
120	0.003	0.543	149	0.023	0.391	192	0.000	0.541

Item	*P-value*	R_{it}	Item	*P-value*	R_{it}
194	0.003	0.610	219	0.035	0.478
195	0.000	0.664	220	0.005	0.499
200	0.005	0.703	221	0.023	0.500
201	0.001	0.614	222	0.014	0.402
202	0.003	0.644	231	0.003	0.543
204	0.009	0.691	233	0.023	0.718
205	0.003	0.433	234	0.005	0.419
206	0.014	0.554	237	0.000	0.647
208	0.000	0.616	239	0.023	0.607
209	0.002	0.752	242	0.009	0.584
210	0.023	0.634	245	0.009	0.612
214	0.001	0.397	246	0.014	0.612
216	0.023	0.389	252	0.001	0.461
217	0.000	0.390	254	0.014	0.495

(VI) The Recalculation of Scores

As mentioned earlier, the purpose of the first stage was to show the difference between holistic scoring system which has been using in majority of universities across the globe and LCI system of scoring to check which one was more objective and consistent:

BY-NC-ND

Table 5. LCI Recalculation of Scores (Par: Participant).

Par	Score	LCI	Par	Score	LCI	Par	Score	LCI	Par	Score	LCI	Par	Score	LCI
1	13	13.863	26	10	8.863	51	13	13.636	76	15	13.863	101	16	17.727
2	15	14.772	27	12	10.227	52	15	12.50	77	15	15.227	102	15	15.909
3	15	16.363	28	13	12.954	53	14	12.045	78	17	17.50	103	17	17.045
4	16	17.727	29	13	13.863	54	10	9.318	79	15	15.454	104	14	14.545
5	14	15.909	30	17	16.818	55	12	10.909	80	16	14.772	105	18	18.409
6	10	8.636	31	10	8.409	56	15	15	81	13	15.681	106	19	17.954
7	11	11.590	32	20	19.090	57	16	17.727	82	19	18.181	107	14	14.090
8	13	12.50	33	20	19.318	58	17	18.181	83	12	11.818	108	15	12.272
9	12	12.272	34	18	19.772	59	17	17.954	84	15	12.50	109	15	13.409
10	17	18.863	35	18	19.545	60	17	17.50	85	16	15.454	110	15	15.227
11	15	14.318	36	9	7.50	61	12	8.863	86	17	15.227	111	18	17.727
12	20	18.409	37	11	8.636	62	12	9.545	87	15	17.727	112	15	14.772
13	18	16.363	38	13	12.045	63	10	10.909	88	14	14.090	113	12	10.909
14	14	12.045	39	14	12.954	64	15	13.181	89	13	15.454	114	10	10.681
15	12	11.818	40	14	16.363	65	15	15.681	90	15	17.50	115	16	17.272
16	11	12.727	41	15	14.545	66	15	16.136	91	14	12.272	116	13	14.772
17	12	12.954	42	12	13.181	67	18	17.045	92	15	13.181	117	14	12.272
18	12	9.318	43	13	15.681	68	18	18.636	93	14	15	118	19	19.090
19	11	10	44	12	9.545	69	20	18.636	94	17	18.409	119	13	10.681
20	15	17.045	45	11	9.090	70	20	19.545	95	12	11.136	120	15	12.954
21	16	17.272	46	15	14.090	71	18	17.954	96	13	11.363	121	20	19.545
22	16	16.136	47	15	15.454	72	17	16.363	97	14	12.954	122	20	19.090
23	17	15.681	48	15	16.136	73	18	17.50	98	15	15.227	123	12	8.409
24	15	16.363	49	15	17.045	74	18	17.045	99	14	15.681	124	16	14.090
25	10	5.227	50	13	15.909	75	18	18.636	100	14	14.090	125	15	14.318

According to Table 5, for instance, the outcome of this recalculation is the most crucial for participants [12] [13], [18], [25], [36], [96], and [123] (just to name a few) going from 20 (holistic scoring) to 18.409, 18 to 16.363, 12 to 9.318, 10 to 5.227, 9 to 7.50, 13 to 11.363, and 12 to 8.409 (LCI scoring) respectively. This is due to the fact that in spite of the overall quality of their translations, they had not been able to translate most of the docimologically calibrated parsing items correctly (88 parsing items) after calculating the degree of item difficulty by 1-PL and r_{it} value (item discrimination). However, for example, participants [34], [35], [50], [87], [89], and [90] (just to name a few) obtained higher scores compared to the first calculation (holistic calculation) (18 *vs.* 19.772), (18 *vs.* 19.545), (13 *vs.* 15.909), (15 *vs.* 17.727), (13 *vs.* 15.454), and (15 *vs.* 17.50) respectively. This was on the grounds that they translated the justified parsing items correctly besides translating the total parses in a text (both justified and unjustified items).

5. DISCUSSION

5.1. The Confrontation between Item Response Theory (IRT) and Classical Test Theory (CTT): The Question of Item Difficulty

Generally, a translation test consists of items which are both easy and difficult to translate. Moreover, some items within a source text discriminate high-performing students from low-performing ones. As explained earlier, stage (III) of LCI method dealt with the measurement of each item's difficulty based on 1-Parameter Logistic Model of IRT. In layman's terms, items difficulty refers to the proportion of correct answers which are provided for one item by the participants. In this direction, higher degree of item difficulty shows the condition when a small percentage of participants get an item correct and conversely lower degree of item difficulty demonstrates the condition that high percentage of participants get an item correct. Both IRT and CTT have been using and testing item difficulty. The question is that why LCI method selects the former approach for measuring item difficulty.

BY-NC-ND

Unlike CTT which ignores the role of participants' competence, IRT is applied to inspect the latent trait (in our case translation competence) which is associated with 'a set of items within a test' (Baker and Kim, 2004). In updated educational contexts, assessment and evaluation are considered inherent parts of curriculum design to evaluate and measure students' proficiency and skill development (Le, 2013). Apart from calculating the total score, a researcher tends to find out whether testing and evaluation tools are adequately formulated to measure a number of specific aspects of students' knowledge. In this direction, the purpose of IRT is to measure/assess/evaluate the relevance of questions coupled with assessing the degree of participants' competence (item difficulty and item discrimination). Unlike CTT, we supposed that a translation participant who translated the source text into plain Persian possessed some amount of translation 'competence' which more or less impacted the end results (see Table 5) (Hambleton, 1989; Kempf, 1983; Farmer *et al.,* 2001; Finch and Edwards, 2015). In order to find out the high-performing translation students from the low-performing ones regarding their translation competence, a number of issues such as the degree of the item difficulty and degree of discriminatory power of each item must be calculated. With that idea, IRT takes into account participants' translation competence and items' characteristics based on item analysis or difficulty (p-value) of 1-Parameter Logistic Model. Unlike CTT, IRT does consider the number of correct items to measure a participant's performance, nor does it suppose 'equal contribution of the items (questions) to the overall scores' (Le, 2013:13). Although items are different in terms of their difficulty and participants differ in terms of their competence or ability, IRT may provide accurate results compared to CTT (Baker, 2001; Zięba, 2013). This is chiefly on the grounds that that participants who obtain the same total scores in a test may vary in their degree of competence. For instance, participants [1], [8], and [28] commonly obtained 13 once the evaluators scored the translations holistically (stage I); however, when the evaluators applied LCI method of translation evaluation, their scores were changed to 13.863, 12.50, and 12.954 respectively. This showed that the participants' level of competence differed with one another in that LCI method could demonstrate such difference. Fox (2010) has pointed out that if a difference between a participant's competence and the level of difficulty of one item is positive or positively skewed (positive reliability coefficient and p-value < 0.05) (e.g. items [99], [134], [185]), the participant has a high chance of answering that item correctly (e.g. participants [40], [81], [94], and [115]). Conversely, if the difference between a participant's competence and the level of difficulty of one item is negative or negatively skewed (negative reliability coefficient and p-value > 0.05) (e.g. items [7], [26], [257]), the participant has the low chance of answering that item correctly (e.g. participants [18], [36], [45], and [69]). As noted earlier, identifying the difficulty of each item is not the absolute condition. Additionally, a translation evaluator must then identify those parsing items which have good and very good discriminatory powers.

5.2. The Confrontation between R_{it} Value and Extreme Group Method: The Question of Item Discrimination

Two ways are available to calculate the item discrimination: (i) the use of r_{it} value (corrected item-total correlation) and (ii) the extreme group method. Translation evaluation models such as PIE method and LCI method depend on discrimination indices to identify higher group of scorers from lower group of scorers. In this direction, in LCI method of evaluation, this stage is computed by the application of r_{it} value function through SAS, SPSS, and Stata statistical packages. A question may arise concerning that why LCI method of evaluation selects the application of r_{it} value to identify high-performing students from the lower ones. As noted earlier, PIE method calculates item discrimination through the application of the extreme group method which dates back to the pre-computer era (Pidgeon and Yates, 1968). Unfortunately, the purpose of using the method of extreme group in PIE method has not so far been explained and substantiated. According to Eyckmans and Anckaert (2017:45),

> The R_{it} value has the advantage over the extreme-group method that every test-taker's score is used to compute the discrimination coefficient, whereas only 54 percent of the test-takers' results are used in the case of the extreme-group method (i.e. the 27% upper and the 27% lower scores).

Kockaert and Segers (2017) as the founder of PIE method believe that the application of 27% rule will maximize differences in a normal distribution; however, it is not the case in point. Selecting the sample size in a normal distribution from a distribution of a test score is of paramount importance. Conventionally, the size of the selected tails is considered an independent sample; nevertheless, the size of the selected tails is dependent and must contain about 21% instead of 27% (D'Agostino and Cureton, 1975). D'Agostino and Cureton (1975:40) have put forward that

> The optimal tail size is around 0.27 if the correlation between the concomitant variable and the test scores is small (i.e. around 0.10). Under normality, this is implying independence. As the correlation increases, the optimal tail size deceases. From above it appears to follow that if the concomitant variable and the test scores have correlation one, then the optimal tail size is around 0.215.

Based on the findings of D'Agostino and Cureton (1975), the data are uncorrelated at the 27% level and yield inaccurate results. The application of r_{it} value has an advantage over the method of extreme group in that it can be used for larger samples. Therefore, an evaluator will have access to the great number of good and very good discriminating items when calculating by r_{it} value. On the other hand, item discrimination which is used in both PIE

BY-NC-ND

method and CTT is determined based on surprisingly small sample size (e.g. from 10 test-takers to 20 test-takers) (Kockaert and Segers, 2014). Consequently, the obtained results must be analyzed and interpreted with great care.

In the context above, after having perused every annotation of the source text and having identified optimal item difficulties based on 1-PL, LCI method selected the application of r_{it} value to identify the top scorers from the bottom ones since this research was dealing with both the large sample size (125 translation students) and the large scale purification, i.e. the process of removing items from multi-item scales due to their negative values or values below 0.30 (see Table 4).

5.3. Why Brat Text Annotator?

As explained earlier, Brat rapid annotation tool is considered an impressionistic web-based tool for making a source text into a number of parses (Bunt *et al.,* 2010). This tool is fully supported by Natural Language Processing (NLP). Brat is taken into account a shared task for detecting how factual statements and parses can be interpreted in terms of their textual contexts involving a hypothesis and an experimental result. The extraction of information into a number of annotations or parses is the fundamental task of representing information contained in a text through the Brat tool. Brat tool is employed to detect metaphor annotation by means of bottom-up identification (Stenetorp *et al.,* 2012) chiefly concentrated on the linguistic metaphors within a source text and extrapolating the conceptual metaphors underlying them. The Brat Stanford CoreNLP tool has a number of features namely; (i) 'high-quality annotation visualization' (every parser is intuitively visualized based on the concept of 'what you see is what you get'); (ii) 'intuitive annotation interface' (this tool is used to detect any parses using Uniform Recourse Identifiers (URLs) that empowers connecting to individual annotations for simplifying easy communication); (iii) 'versatile annotation support' (Brat is set up to support most annotation tasks by means of binary relations such as part of speech (POS) tagged tokens or chunks); and (iv) 'corpus search functionality' (this tool executes an exhaustive set of search functions which permits users to search through the collection of different documents for text span relations and correspondents) (Stenetorp *et al.,* 2012). For example, in Figure 2, every annotation within a source text is visually linked to different colors. Moreover, the Brat tool detects relations between tokens or chunks so that an evaluator can simply spot the corresponding parsing items in a reciprocal language to scrutinize whether or not the detected parsing item is correctly translated.

LCI moves from the norm-referenced assessment to the criterion-referenced assessment. Unlike available translation evaluation methods such as PIE method, analytic method, etc. which focus on either criterion-referenced assessment or norm-referenced assessment, LCI method benefits from the amalgam of norm and criterion-referenced assessment methods by means of a feedback loop, including a norm-referenced assessment method (the whole parsing items in a sauce text), criterion-referenced assessment (the docimologically justified parsing items), and the actual evaluating. This feedback loop is used to remove any score inflation by the concomitant use of both norm- and criterion-referenced assessment methods.

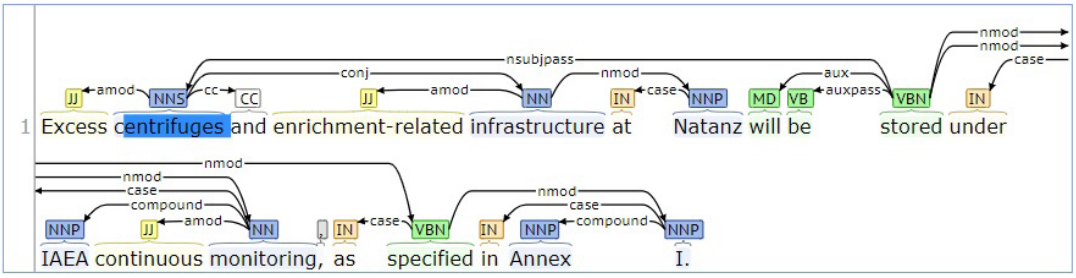

Figure 2. Illustration of Brat Software Analysis.

For example, the term 'infrastructure' (*NN*) within the source text relates to the corresponding terms such as 'enrichment-related' (*JJ-amod-NN*) and 'centrifuges' (*NNS-conj-NN*) on the left side and 'Natanz' (*NN-nmod-NNP*) on the right side. In this vein, evaluators must search for the corresponding translations of the terms 'enrichment-related-***infrastructure***', 'centrifuges- ***infrastructure***', and '***infrastructure***-Natanz'. Based on the four representative translations, the corresponding Persian translations are '*zir sāxt'hāje qæni sāzi'je mærbouteh*', '*sāntrifiouzh'ha va zir sāxt*', and '*zir sāxt dar Nætænz*' respectively. Then, the evaluators compare students' translations with the given corresponding translations to check their acceptability.

To take another example, the term 'stored' (*VBN*) corresponds to 'centrifuge-**stored**' (*NNS-nsubjpass-VBN*), 'will-**stored**' (*MD-aux-VBN*), and 'be-**stored**' (*VB-auxpass-VBN*) on the left side and '**stored**-monitoring' (*VBN-nmod-NN*) and '**stored**-specified' (*VBN-nmod-VBN*) on the right side. The corresponding translations based on representative translations are '*sāntrifiouzh'haje ænbār šode*', '*ænbār kærdæn*', '*ænbār xāhæd šod*', '*tæhte nezāræt ænbār šode*' and '*ænbār šode be næhve mæš'ruh*' respectively. These translations were approved by the

BY-NC-ND

evaluators as correct corresponding translations. Last but not least, based on the extracted chunks, evaluators measured the acceptability of the translations (optimal p and r_{it} values) to check whether or not a parsing item is tagged as a docimologically justified item.

6. CONCLUSION

6.1. Limitations of the Research

One of the limitations of the present research is the relatively small sample size. As far as 1-PL of IRT is used for a larger sample size (Hambleton and Jones, 1993:43), the maintained findings of the present research paper are not adequately accurate when compared with a large sample size (above 500 participants). Also, the translation assignment (English to Persian translation) was carried out with paper and pen. In a recreation of the research paper with a larger sample size, much prudence is needed to permit the involved participants to conduct the assignment by computer or online-platforms. As far as the administration of 1-PL is manually time-consuming; thus, a good knowledge of statistical packages (e.g. Stata, SAS, and Winsteps) are pivotally needed for a researcher to precisely and meaningfully interpret and analyze the obtained results. Another limitation is that the application of LCI method is a time-consuming activity. To solve such a problem, a computerized platform is necessarily required to check responses and likewise a list of correct and incorrect solutions of annotations must be arranged.

6.2. Implications of the Research

It is widely known that revising and scoring translations are time-consuming, tedious, repetitive, and subjective. A reviser cannot guarantee that he/she can find every mistake in translations of the same source text and give consistent feedback to translators or grade translations consistently. These are the challenges that translationQ solves in the field of translation evaluation and assessment. TranslationQ is a web and cloud-based productivity software that a translator can use for translation training and revision. This online platform ensures objectivity and works independent of language or domain. No matter which language pair is being used and no matter if a translator is using different text types. When a translator prepares a translation, the end product can be imported in the formats of SDL XLIFF and word files. TranslationQ segments a source text automatically. Also, translationQ tags errors for bilingual text through using TAUS Dynamic Quality Framework (DQF) error categories 'that was developed as part of the (EU-funded) QTLaunchPad project (large-scale action for quality translation technology) based on careful examination and extension of existing quality models' (TAUS, 2018). TAUS DQF refers to 'a comprehensive set of tools, best practices, metrics, reports and data to help the industry set benchmarking standards' (ibid.).

TranslationQ will automatically detect identical errors in other translations and this will save a huge amount of time. Blind revision of translations is another possibility in case where researchers and instructors tend to use translationQ for high-stake exams. In this light, LCI method has the potentiality to be employed in the translationQ platform, because this method can be used in various text types such as medical, political, cultural, and so forth. As far as the application of LCI method is a laborious activity, the automation of this method by the translationQ will improve the scoring system, add options throughout the Brat process (Stage II), and update all available corrections regularly.

In summary, this paper introduced a translation evaluation method called Logistic Calibrated Items (LCI). The purpose of LCI method was to objectify translation evaluation products. LCI method was an attempt to identify high proficient translators through the application of six phases as fully explained in section 4. Last but not least, LCI employed Brat CoreNLP software to identify all parses (norm-referenced assessment method) and then distinguish appropriate and justified items (optimal p and r_{it} values) (criterion-referenced assessment method) within a source text.

REFERENCES

Akbari, A. & Gholamzadeh Bazarbash, M. (2017). "Holistic Assessment: Effective or Lenient in Translation Evaluation?" *Skopos. Revista Internacional de Traducción e Interpretación, 8*/1: 51-67.

Akbari, A. & Segers, W. (2017a). "Translation Difficulty: How to Measure and What to Measure". *Lebende Sprachen, 62*/1: 3-29. https://doi.org/10.1515/les-2017-0002

Akbari, A. & Segers, W. (2017b). "Translation Evaluation Methods and the End-Product: Which One Paves the Way for a More Reliable and Objective Assessment?" *Skase Journal of Translation and Interpretation, 11*/1: 2-24.

Akbari, A. & Segers, W. (2017c). "Evaluation of Translation through the Proposal of Error Typology: An Explanatory Attempt". *Lebende Sprachen, 62*/2: 408-430. https://doi.org/10.1515/les-2017-0022

BY-NC-ND

Akbari, A. & Segers, W. (2017d). "The Perks of Norm and Criterion Referenced Translation Evaluation". LICTRA, Leipzig, Germany, 20 March.

Anckaert, P., Eyckmans, J. & Segers, W. (2008). "Pour Une Évaluation Normative De La Compétence De Traduction." *ITL - International Journal of Applied Linguistics, 155*/1: 53-76. https://doi.org/10.2143/ITL.155.0.2032361

Bahameed, A. S. (2016). "Applying assessment holistic method to the translation exam in Yemen." *Babel, 62*/1: 135-149. https://doi.org/10.1075/babel.62.1.08bah

Baker, F. B. (2001). *The Basics of Item Response Theory*. 2nd ed. New York: ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B. & Seock-Ho K. (2004). *Item Response Theory: Parameter Estimation Techniques*. 2nd ed. New York: Marcel Dekker. https://doi.org/10.1201/9781482276725

Barkaoui, K. (2010). "Explaining ESL essay holistic scores: A multilevel modeling approach". *Language Testing, 27*/4: 515-535. https://doi.org/10.1177/0265532210368717

Barkaoui, K. (2011). "Effects of marking method and rater experience on ESL essay scores and rater performance". *Assessment in Education: Principles, Policy & Practice, 18*/3: 279-293. https://doi.org/10.1080/0969594X.2010.526585

Bunt, H., Merlo, P. & Nivre, J. (2010). *Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing*. Netherlands: Springer. https://doi.org/10.1007/978-90-481-9352-3

Conde Ruano, T. (2005). "No me parece mal. Comportamiento y resultados de estudiantes al evaluar traducciones". Unpublished doctoral dissertation. University of Granada, Granada.

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

D'Agostino, R. B. & Cureton, E. E. (1975). "The 27 Percent Rule Revisited". *Educational and Psychological Measurement, 35*/1: 47-50. https://doi.org/10.1177/001316447503500105

Dancette, J. (1989). "La faute de sens en traduction". *TTR : traduction, terminologie, rédaction, 2*/2: 83-102. https://doi.org/10.7202/037048ar

Eberly Center. (2016). "What is the difference between formative and summative assessment?" Accesible at https://www.cmu.edu/teaching/assessment/basics/formative-summative.html [Last access: July 2019].

Eyckmans, J. & Anckaert, P. (2017). "Item-based assessment of translation competence: Chimera of objectivity versus prospect of reliable measurement". *Linguistica Antverpiensia, New Series: Themes in Translation Studies, 16*/1: 40–56.

Eyckmans, J., Anckaert, P. & Segers, W. (2013). "Assessing Translation Competence". *Actualizaciones en Comunicación Social. Centro de Lingüística Aplicada, Santiago de Cuba, 2*, 513-515.

Eyckmans, J., Segers, W. & Anckaert, P. (2012). Translation Assessment Methodology and the Prospects of European Collaboration. In *Collaboration in Language Testing and Assessment*, edited by D. Tsagari and I. Csépes, 171-184. Bruxelles: Peter Lang.

Farmer, W. L., Thompson, R. C., Heil, S. K. R. & Heil, M. C. (2001). Latent Trait Theory Analysis of Changes in Item Response Anchors. Accesible at https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/2000s/media/0104.pdf. [Last access: July 2019].

Finch, H. & Edwards, J. M. (2015). "Rasch Model Parameter Estimation in the Presence of a Nonnormal Latent Trait Using a Nonparametric Bayesian Approach". *Educational and Psychological Measurement, 76*/4: 662-684. https://doi.org/10.1177/0013164415608418

Fox, J. (2010). *Bayesian Item Response Modeling: Theory and Applications*. Amsterdam: Springer. https://doi.org/10.1007/978-1-4419-0742-4

Garant, M. (2009). "A case for holistic translation assessment". *AFinLA-e: Soveltavan kielitieteen tutkimuksia, 5*/2: 5-17.

Gonzalez, K. (2018). "Contrast Effect: Definition & Example". https://study.com/academy/lesson/contrast-effect-definition-example.html.

Gouadec, D. (1989). "Comprendre, évaluer, prévenir : Pratique, enseignement et recherche face à l'erreur et à la faute en traduction". *TTR, 2*/2: 35-54. https://doi.org/10.7202/037045ar

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In *The American Council on Education/Macmillan series on higher education*, edited by R. L. Linn, 147-200. New York, NY, England: Macmillan Publishing Co.

Hambleton, R. K. & Jones, R. W. (1993). "Comparison of classical test theory and item response theory and their applications to test development". *Educational Measurement: Issues and Practice, 12*/3: 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In *Assessing second language writing in academic contexts*, edited by L. Hamp-Lyons, 241–76. Norwood, NJ: Ablex.

BY-NC-ND

Harsch, C. & Martin, G. (2013). "Comparing holistic and analytic scoring methods: issues of validity and reliability". *Assessment in Education: Principles, Policy & Practice, 20*/3: 281-307. https://doi.org/10.1080/0969594X.2012.742422

Kempf, W. (1983). "Some Theoretical Concerns ahout Applying Latent Trait Models in Educational Testing". Accesible at https://pdfs.semanticscholar.org/5909/0351c0bc109f28836a75eaa67e7eecaffa41.pdf. [Last access: July 2019].

Kockaert, H. & Segers, W. (2014). "Evaluation de la Traduction : La Méthode PIE (Preselected Items Evaluation)". *Turjuman, 23*/2: 232-250.

Kockaert, H. & Segers, W. (2017). "Evaluation of legal translations: PIE method (Preselected Items Evaluation)". *Journal of Specialized Translation, 27*: 148-163.

Kussmaul, P. (1995). *Training the Translator*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/btl.10

Larose, R. (1989). *Théories contemporaines de la traduction*. Montréal: Presses de l'Université de Québec.

Le, D. (2013). Applying item response theory modeling in educational research. Instructional Technology, Graduate College at Iowa State University Digital Repository.

Lei, P. & Wu, Q. (2007). "CTTITEM: SAS macro and SPSS syntax for classical item analysis". *Behavior Research Methods, 39*/3: 527-530. https://doi.org/10.3758/BF03193021

Mariana, V., Cox, T. & Melby, A. (2015). "The Multidimensional Quality Metrics (MQM) Framework: A New Framework for Translation Quality Assessment". *Journal of Specialized Translation, 23*:137-161.

Muñoz Martín, R. (2010). On Paradigms and Cognitive Translatology. In *Translation and Cognition*, edited by G. Schreve and E. Angelone, 169-187. Amsterdam and Philadelphia: John Benhamins. https://doi.org/10.1075/ata.xv.10mun

Newmark, P. (1991). *About Translation*. Clevedon: Multilingual Matters.

Nord, C. (2005). *Text Analysis in Translation*. aMSTERDAM: Rodopi.

Pidgeon, D. A. & Yates, A. (1968). *An introduction to educational measurement*. London: Routledge.

Preacher, K. J. , Rucker, D. D., MacCallum, R. C. & Nicewander, W. A. (2005). "Use of the extreme groups approach: a critical reexamination and new recommendations". *Psycholological Methods, 10*/2: 178-792. https://doi.org/10.1037/1082-989X.10.2.178

Schmitt, P. A. (2005). Qualitätsbeurteilung von Fachübersetzungen in der Übersetzerausbildung, Probleme und Methoden. Vertaaldagen Hoger Instituut voor Vertalers en Tolken, 16-17 March.

SPSS, IBM. (2017). Available at https://www.ibm.com/analytics/us/en/technology/spss/. [Last access: July 2019].

Stansfield, C. W., Scott, M. L. & Kenyon, D. M. (1992). "The Measurement of Translation Ability". *The Modern Language Journal, 76*/4: 455-467. https://doi.org/10.2307/330046

Stata. (2016). "Stata: Software for Statistics and Data Science". Available at https://www.stata.com/. [Last access: July 2019].

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S. & Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France.

Stenetorp, P., Topic, G., Pyysalo, S., Ohta, T., Kim, J. & Tsujii, J. (2011). BioNLP Shared Task 2011: Supporting Resources. BioNLP Shared Task 2011 Workshop, Portland, Oregon, USA.

TAUS. (2018). "Measuring and Benchmarking Translation Quality". Available at https://www.taus.net/quality-dashboard-lp. [Last access: July 2019].

Umobong, M. E. (2017). "The One-Parameter Logistic Model (1PLM) And its Application in Test Development". *Advances in Social Sciences Research Journal, 4*/24: 126-137.

Van Antwerpen, J. (2016). "P-, D-, and Rit values: a new start". Available at http://www.andriesseninternational.com/p-d-and-rit-values-a-new-start/. [Last access: July 2019].

Waddington, C. (2001). "Different Methods of Evaluating Student Translations: The Question of Validity". *Meta, 46*/2: 311-325. https://doi.org/10.7202/004583ar

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Widdowson, H. G. (1978). *Teaching Language as Communication* Oxford: Oxford University Press.

Wiersma, W. & Jurs, S. G. (1990). *Educational Measurement and Testing*. London: Allyn and Bacon.

Zięba, A. (2013). "The Item Information Function in One and Two-Parameter Logistic Models- A Comparison and Use in the Analysis of the Results of School Tests". *Didactics of Mathematics, 10*/14: 87-96. https://doi.org/10.15611/dm.2013.10.08

BY-NC-ND

APPENDIX 1

Selected Source Text

Iran's long term plan includes certain agreed limitations on all uranium enrichment and uranium enrichment-related activities including certain limitations on specific research and development (R&D) activities for the first 8 years, to be followed by gradual evolution, at a reasonable pace, to the next stage of its enrichment activities for exclusively peaceful purposes, as described in Annex I. Iran will abide by its voluntary commitments, as expressed in its own long-term enrichment and enrichment R&D plan to be submitted as part of the initial declaration for the Additional Protocol to Iran's Safeguards Agreement. Iran will begin phasing out its IR-1 centrifuges in 10 years. During this period, Iran will keep its enrichment capacity at Natanz at up to a total installed uranium enrichment capacity of 5060 IR-1 centrifuges. Excess centrifuges and enrichment-related infrastructure at Natanz will be stored under IAEA continuous monitoring, as specified in Annex I. Iran will continue to conduct enrichment R&D in a manner that does not accumulate enriched uranium. Iran's enrichment R&D with uranium for 10 years will only include IR-4, IR-5, IR-6 and IR-8 centrifuges as laid out in Annex I, and Iran will not engage in other isotope separation technologies for enrichment of uranium as specified in Annex I. Iran will continue testing IR-6 and IR-8 centrifuges, and will commence testing of up to 30 IR-6 and IR-8 centrifuges after eight and a half years, as detailed in Annex I.

APPENDIX 2: WADDINGTON'S FRAMEWORK OF HOLISTIC TRANSLATION

Level	Accuracy of Transfer of ST Content	Quality of Expressions in TL	Degree of Task Completion	Mark
Level 5	Complete transfer of ST information, only minor revision needed to reach professional standards.	Almost all the translation reads like a piece originally written in English. There may be minor lexical, grammatical, and spelling errors.	Successful	9,10
Level 4	Almost complete transfer; there may be one or two insignificant inaccuracies; requires certain amount of revision to reach professional standards.	Large sections read like a piece originally written in English. There are a number of lexical, grammatical, or spelling errors.	Almost completely successful	7,8
Level 3	Transfer of general ideas but with a number of lapses in accuracy; needs considerable revision to reach professional standards.	Certain parts read like a piece originally written in English, but others read like a translation. There are a considerable number of lexical, grammatical, or spelling errors.	Adequate	5,6
Level 2	Transfer undermined by serious inaccuracies; thorough revision required to reach professional standards.	Almost the entire text reads like a translation; there are continual lexical, grammatical, or spelling errors.	Inadequate	3,4
Level 1	Totally inadequate transfer of ST content, the translation is not worth revising.	The candidate reveals a total lack of ability to express himself adequately in English.	Totally inadequate	1,2

BY-NC-ND

BY-NC-ND