Translation Evaluation Methods and the End-Product: Which One Paves the Way for a More Reliable and Objective Assessment?

Alireza Akbari, Winibert Segers

Abstract
This research paper tried to introduce a method for evaluating the end-products known as Preselected Items Evaluation (PIE) method devised by Hendrik Kockaert and Winibert Segers. This method maximized students’ performance by means of p-docimology (item difficulty) and d-index (Item discrimination). This paper presented the results of a controlled empirical design in which three methods of translation evaluation such as holistic, analytic, and PIE were juxtaposed in terms of inter-rater reliability and correlation coefficient. The results showed that in spite of the fact that holistic and analytic methods lacked test reliability, PIE method compensated test validity and reliability through distinguishing some preselection of items having good p-values and d-indices.

Keywords: Holistic Method, Analytic Method, Quality of Translation, Preselected Items Evaluation (PIE) Method, P-docimology, D-index

Introduction

Recently, the need for more practical and empirical evidences for the quality of translation tests in the field of translation studies has been professed (Anckaert and Eyckmans 2006; Anckaert et al. 2008). Although professional and educational institutions have organized translation skills workshops in terms of test administrations, the validity and the reliability of those tests are undetermined (Eyckmans et al. 2009). With this in mind, evaluating the quality of translation is heavily relied upon the ‘code of practice’ rather than ‘empirical research’ (ibid.). The term translation evaluation aims at inspecting and marking the students’ end-products (translations) in terms of certain translation theory. Comparison of the source text with the target one and also the target text per se are the processes of translation evaluation (Mobaraki and Aminzadeh 2012). According to Rebert Lado (1961), ‘translation tests that are so common in testing language proficiency skills are not available as tests of the ability to translate’. To the best of our knowledge, there does not exist any method to scrutinize translation competence psychometrically (Anckaert et al. 2008). To prove this, there avail two reasons: (1) the lack of translation tests validity and (2) the epistemological aperture confronted between the protagonists of hard sciences versus humanities (Eyckmans et al. 2009). Through evaluating translation, the aspects such as linguistic, paralinguistic, translation process, situatedness (Muñoz Martín 2010), strategies (e.g. power relations, pragmatics, and discourse), translation service, and finally translator’s competence (e.g. cultural, technical, and linguistic competences) (Kockaert and Segers 2016) are applied. Practice of evaluating translation is carried out on the basis of criterion referenced approach in both educational and professional contexts (Al-Qina 2000; Schmitt 2005). These contexts are evaluated in terms of some assessment grids to make the translation evaluation more reliable and valid. Although these assessment grids are based on the evaluator’s intention to determine various dimensions of translation competence, they are unable in decreasing the
subjectivity of translation evaluation sufficiently. This is due to the fact that the recognition of ‘dimensions of translation competence’ is chiefly subjective in nature (Anckaert et al. 2008). Other factor which threatens the reliability of a translation test is the system of scoring susceptible to contrast effects such as halo effect. It is so demanding to provide a good test in which one evaluator does not only determine good translations from bad ones, but also he/she can distinguish the average quality of translations.

The Purpose of the Research

The present research is an attempt at evaluating the inter-rater reliability and correlation coefficient of three methods of assessment such as holistic, analytic, and Preselected Items Evaluation (PIE) methods on the basis of their psychometric qualities in a controlled empirical design. This research will gather the obtained data within an educational context. But, the insights concerning translation competence assessment can be operated in professional setting.

The Objectives of the Research

The research objectives of the present research are as follows:

1. To calculate the interrater reliability correlation for holistic, analytic, and PIE methods;
2. To calculate the correlation coefficient between holistic, analytic, PIE methods to see their strengths and weaknesses when evaluating a translation product.

Research Hypotheses

The fundamental research question of the present research concentrates on the reliability of the mentioned assessment methods with reference to one another. Inter-rater reliability correlation alludes to the consistency of the methods and the results. In translation evaluation, inter-rater reliability refers to when a translation product is evaluated similarly by various evaluators utilizing the similar method of evaluation. In this direction, the reliability of holistic and analytic methods is indirectly assessed through Spearman rho rank correlation coefficient of different evaluators’ scores. However, the inter-rater reliability of PIE method is calculated, firstly by means of ‘docimological p (item difficulty)’ and ‘IDis (item discrimination) (by means of the rule of 27 percent)’. Then the data obtained by p-value and d-index is imported to SPSS (version.24) to see the amount level of agreement between the raters through Cronbach’s Alpha. With this in mind, the main research question is divided into several hypotheses perusing as follows:

Inter-rater Reliability Hypotheses

1. The inter-rater reliability of holistic method is higher than the inter-rater reliability of PIE method;
2. The inter-rater reliability between the evaluators will be in between in the case of analytic method;
3. And due to the subjective results of both holistic and analytic methods, they will not be able to discriminate students’ performance. In case of PIE method, the discriminatory power is maximized through the preselected items in terms of ‘p-docimology’ and ‘d-index’.

**Correlation Strength Hypotheses**

1. The correlation between the holistic method and the analytic method is strong;
2. The correlation between the holistic method and the PIE method is strong;
3. The correlation between the analytic method and the PIE method is strong.

**State of the Art**

**Translation Assessment**

Researchers from the domain of translation assessment have declared that this field is still in its infancy (Waddington 2001; Gile 2005). As a matter of fact, no association of evaluating translation has so far seen the light and no conferences has been held particularly aimed at this issue. Generally, evaluating and assessing translation skills are in the hands of practitioners rather than researchers. Therefore, it is hardly surprising that translation evaluation is more relied upon the ‘code of practice’ rather than on empirical and practical researches (Eyckmans et al. 2012). Most of the state of the arts on translation evaluation and assessment are about the role of error detection, error distinctions in terms of their significance (Conde Ruano 2006), and the locations of errors within a text (Vollmar 2001). Within the past 25 years, numerous models and assessment grids have been proposed for a variety of various texts and genres to contribute to the objectivity of evaluating translation (Al-Qinai 2000; Schmitt 2005).

In the articles carried out by Anckaert et al. (2008) and Eyckmans et al. (2009), they inspect the ‘methodology of educational measurement’ as well as language testing theory insights to the field of translation studies. They aim at setting up translation assessment practices which are more evaluator-independent. They have developed a norm referenced method to evaluate translation skills (Eyckmans et al. 2012) to move away from the criterion-referenced analytical method. According to Eyckmans et al. (2012:177),

The method hinges on the assumption that translation assessment should be based on a consideration of the performance on the entire text, since in actual performances subcomponents of translation ability are more or less inextricable and mistakes as well as bonuses originate from the interaction of a particular text with a particular translator.

This approach which is called Calibration of Dichotomous Items (CDI) consists of conveying the items to the domain of translation assessment and studies. With this in mind, the performance of a group of test takers is utilized to distinguish the text segments having discriminating power. They also argued that ‘every element of the text that contributes to the measurement of differences in translation ability acquires the ‘item’-status’. Finally, this
method calculates item difficulty, discrimination indices, and the reliability of text segments on the basis of Cronbach’s Alpha.

**Translation Evaluation Methods: Review**

**Holistic Method**

Holistic method evaluates the quality of translation on the basis of translator’s overall impressions (intuition) (Mariana et al. 2015; Eyckmans et al. 2009) and applies in professional setting. In order to evaluate a translation in terms of holistic method, the grader or evaluator scores the intended translation through his/her overall intuition as excellent, good, fair, or very bad. Therefore, on can conclude that this method is fully-fledged subjective. According to Kockaert and Segers (2016), ‘holistic approach seemed to focus better on a context sensitive evaluation, and seemed to discard exclusive attention to grammatical errors in the translation tests’. In other word, ‘translations do not occur in a vacuum that they need to be interpreted and evaluated in their relevant context’ (Koskinen 2008). In this direction, care must be taken not to label holistic method as an unsystematic method, since Garant (2009:10) argues that

Holistic method refers to a systematic way in which the teacher arrives at an overall impression of the text as opposed to relying on a discrete point-based scale. The teachers in that group had each devised their own, systematic way of evaluating translations.

Waddington (2001) describes four methods of assessment in which the third one is associated with the role of holistic assessment in translation evaluation. He designed a survey and the teacher were asked to send a brief description of the methods of assessment they applied during their class hour. In the upshot, he received only three descriptions of holistic assessment. Then he designed the following paradigm on holistic method of assessment (scores from 0 to 10):

*Table 1. Holistic Method (Adopted from Waddington 2001:214)*

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

**Analytic Method**

Analytic assessment is a method to evaluate translation quality through inspecting the text segment (e.g. sentence, paragraph, individual words, etc.) in terms of certain criteria rather than concentrating on the evaluator’s overall impression. According to Waddington
(2001:36), analytic assessment which is based on error analysis is ‘more reliable and valid than holistic method’. Eyckmans et al. (2013) argue that translation errors must be numbered on the basis of the evaluation grid criteria provided below. With this idea, the evaluator must underline every error such as language and translational errors and provide the pertinent information in the margin due to the nature of the errors.

**Table 2. Analytic Method (Adopted from Eyckmans et al. 2013)**

<table>
<thead>
<tr>
<th>Description</th>
<th>Example</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SENS</strong> (Meaning or Sense)</td>
<td>Toute altération du sens dénotatif: informations erronées, nonsense, important omission...</td>
<td>-1</td>
</tr>
<tr>
<td><strong>CONTRESENS</strong> (Misinterpretation)</td>
<td>L'étudiant affirme le contraire de ce que dit le texte: information présentée de manière positive alors qu'elle est négative...</td>
<td>-2</td>
</tr>
<tr>
<td><strong>VOCABULAIRE</strong> (Vocabulary)</td>
<td>Choix lexical inadapté, collocation insitée... (Unsuited lexical choice, use of non-idiomatic collocations)</td>
<td>-1</td>
</tr>
<tr>
<td><strong>CALQUE</strong> (Calque)</td>
<td>Utilisation d’une structure littéralement copiée et insitée en français... (Cases of literal translation of structures, rendering the text into French)</td>
<td>-1</td>
</tr>
<tr>
<td><strong>REGISTRE</strong> (Register)</td>
<td>Selon la nature de texte ou la nature d’un extrait (par exemple, un dialogue): traduction trop (in)formelle... (Translation that is too (in)formal or simplistic and not corresponding to the nature of the text or extract)</td>
<td>-0.5</td>
</tr>
<tr>
<td><strong>STYLE</strong> (Style)</td>
<td>Lourdeurs, répétition maladroits, assonances... (Awkward tone, repetition, unsuited assonances)</td>
<td>-0.5</td>
</tr>
<tr>
<td><strong>GRAMMAIRE</strong> (Grammar)</td>
<td>Erreurs grammaticales en français... (par exemple, mauvais accord du participe passé, confusion masculine/féminin, accords fautifs...) + mauvaise compréhension de la</td>
<td>-0.5</td>
</tr>
</tbody>
</table>
grammaire du texte original (par exemple, un passé rendu par un présent…) et pour autant que ces erreurs ne modifient pas en profondeur le sens.

(Grammatical errors in French (for example, wrong agreement of the past participle, gender confusion, wrong agreement of adjective and noun, ....) + faulty comprehension of the grammar of the original text (for example, a past event rendered by a present tense, ....), provided that these errors do not modify the in-depth meaning of the text)

| OUBLIS (Omission) | Voir SENS (See sense/ meaning) | -1 |
| AJOUTS (Addition) | Ajout d’informations non contenues dans le texte (sont exclus de point les étofements stylistique) | -1 |
| ORTHOGRAPHE (Spelling) | Erreurs orthographiques, pour autant qu’elles ne modifient pas le sens. | -0.5 |
| PONCTUATION (Punctuation) | Oubli ou utilisation fautive de la ponctuation. Attention: l’oublì, par exemple, d’une virgule induisant une compréhension différente du texte, est considéré comme une erreur de sens. | -0.5 |

Compared to the holistic assessment, analytic assessment nictitates allotting more time to the ins and outs of the text. Therefore, the translator has a better insight into the correctness of translation in analytic method. To cut a long story short, analytic assessment is also subjective, since various evaluators and graders are not always in the position of agreement. One minor error for one grader is considered a major error for another grader.

Preselected Items Evaluation (PIE) Method
The preselected items evaluation (PIE) method is the feasible version of Calibration of Dichotomous Items (CDI) method (see previous section) in translation evaluation (Anckaert et al. 2013; Kockaert and Segers 2012; 2014). PIE method is a system designed by Hendrik Kockaert and Winibert Segers highly appropriate for summative assessment. This method is on the basis of the number of preselected items constrained in evaluation. The intended assessment method is also calibrated and dichotomous method covering both norm and criterion referenced assessments. With this idea, the preselected items in the source language are selected in terms of their item difficulty (docimological p) and item discrimination (d-index) for which the 'correct and erroneous solutions are determined’ (Kockaert and Segers 2016). In the light of these explanations, one has to become familiar with the two terms such as item difficulty and item discrimination:

- **Item difficulty (Docimological-p)**

  Rasinger (2008:159) argues that ‘significance in statistics refers to the probability of our results being a fluke or not; it shows the likelihood that our result is reliable and has not just occurred through the bizarre constellation of individual numbers.’ Moreover, statistical significance can be substantiated through ‘docimological-p’ illustrating the 'probability that the results happened by chance' (Saldanha and O'Brien 2013:199). In other word, docimological-p can be described as the ratio of participants answering an item correctly with the total examination participants:

  \[ P_i = \frac{R_i}{N_i} \]

  \( P_i \) refers to docimological-p or the difficulty factor of preselected item, \( R_i \) alludes to a number of participants answer an item correctly, and \( N_i \) refers to the total population of participants. Both PIE and CDI methods involve the calculation of item difficulty and item discrimination for every item in CDI and for the number of preselected items in PIE method.

- **Item Discrimination (d-index)**

  Ary et al. (2010:211) express that ‘the item discrimination index shows the extent to which each item discriminates among the respondents in the same way as the total score discriminates.’ Item discrimination can be carried out through correlating item scores with total scale scores. If the high scorers on an individual item have the high total scores and if the low scorers have the low total scores, therefore, the item is discriminating in the same way as the total score. In order that the item discrimination is to be useful, the item must correlate at least 0.25 with the total score. Items having low correlation or negative correlation with the total score should be discarded owing to the fact that they are not contributing to the measurement of the attitude (ibid). Positive item discrimination is productive unless ‘it is so high that the item merely repeats the information provided by other items on the test’ which is called the 'attenuation paradox' (Ebel 2002). Ebel (2002) argues that:

  The discrimination index (d) is computed from equal sized high and low scoring groups on the test. Subtract the number of successes by the low group on the item from the number of successes by the high group, and
then divide this difference by the size of the group. The range of this index is +1 to -1. Using Truman Kelley’s ‘27% of sample’ group size, values of 0.40 and above are regarded as higher and lesser than 0.20.

Wiersma and Jurs (1990:145) note that ‘27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis.’

Table 3. Comparison of All Assessment Methods (Adopted from Kockaert and Segers 2016)

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Holistic</th>
<th>Analytic</th>
<th>CDI</th>
<th>PIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Items</td>
<td>Exhaustive</td>
<td>Exhaustive</td>
<td>Docimologically Relevant Items</td>
<td>Translation Brief</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>Global</td>
<td>Grids/Criteria</td>
<td>Grids/Criteria</td>
<td>Grids/Criteria</td>
</tr>
<tr>
<td>Acceptance of Alternatives</td>
<td>Expected/Unexpected</td>
<td>Expected/Unexpected</td>
<td>Expected/Unexpected</td>
<td>Expected/Unexpected</td>
</tr>
<tr>
<td>EN 15038 Compatible</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Interrater Reliability</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Criterion Referenced</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Norm Referenced</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Methodology

Subjects

A total number of 40 translator students participated in this study. They were enrolled in the second year master-level English-Persian translation workshops and courses from the University of Isfahan. The participants’ ages range from 22-25. The translation assignment was carried out in class hour under the supervision of the researcher and the translation trainer.

Corpus of the Research
The participants in this research were asked to translate an excerpt from ‘Joint Comprehensive Plan of Action’ (The International Agreement on the Nuclear Program of Iran) between Iran and P5+1 (China, United Kingdom, United States, Russia, France, and Germany) into Persian. This text elucidates the ‘Preamble and General Provision’ of the nuclear agreement. The reasons to choose such political text are mainly due to its technical language illustrating the socially significant phenomena and its terminologies. According to Huseynova (2015:150), ‘vocabulary [terminology] of political texts has a heterogeneous composition’. The most common layers are the general lexica-political and syntactic-political areas. As such, the length, register, type, and source language difficulty can be regarded expressive of the materials covered in the translation courses. Also, this research applies ten different translations to pave the way for the evaluators/raters having the spectrum of correct equivalents of the source text while scoring the translations.

Materials and Procedures

The translator students of the University of Isfahan were asked to translate a relatively short text (236 words) from English into Persian (L2 to L1). The translation drafts were handed to 6 raters asked to correct them using holistic method (2 raters), 2 raters who were asked to correct the translations using analytic method, and 2 raters asked to correct them utilizing PIE method. The evaluators were selected on the basis of their longstanding experience in translation evaluation. The evaluators adopted for the analytic and holistic methods were asked to score all the translation drafts on the basis of Eyckmans et al. (2013) and Waddington’s (2001) frameworks (out of 20) respectively. The evaluators using PIE method scored the translation drafts through docimological-p and d-index. The internal consistency of PIE method was calculated through Cronbach’s Alpha. The students were given 90 minutes to render the text. Also, students were allowed to access their computers to use whatever electronic references, since this research paper tried to make sure that heuristic competence was included in the translation performance. And in the meantime, the students were not allowed to consult with one another during the assignment. They were given instructions in their mother tongue (Persian) and the collection of data was supervised by the translator trainer and the researcher.

Processing of Data

The raters/evaluators were informed about the quasi-experimental design of the study before they showed their agreement to participate in this study. Tellingly, with their participation, they received an appropriate payment for the amount of time allotted to score all the translations. However, identifying and selecting the raters are also important. This research approached these raters according to their profiles. Also, this research paved the way for the raters to choose whatever methods they felt comfortable preferred to work with (holistic, analytic, and PIE methods). For statistical reasons, this yielded in a selection of 2 raters for holistic method, 2 raters for analytic method, and 2 raters for PIE method having experienced translation training with more than 15 years of experience in evaluating students’ translation performance. Therefore, the profiles of the raters taking part in this research were as follows:

- **Raters in Holistic Method**
Rater (1) was a translator trainer utilizing holistic method throughout his career when assessing students’ performance. He was convinced of the advantages of holistic assessment. Rater (2) was a professional translator approaching translation assessment holistically and he scrutinized the points made by rater (1).

- **Raters in Analytic Method**

Rater (1) was a translator trainer having been using analytic method throughout his career. His scoring sheets were examined and approved by Rater (2). Rater (2) was a professional translator working with analytic method. He was responsible for approving the assessment sheets of Rater (1).

- **Raters in PIE Method**

Rater (1) was a translator trainer having been using PIE method throughout his career. His scoring assessments were examined and approved by Rater (2) going to objectively validate students’ translations of the preselected items through item difficulty (p-docimology) and discrimination index (d-index). Rater (2) was a professional translator working with PIE method. He was responsible for approving the evaluations of Rater (1) through item difficulty (p-docimology) and discrimination index (d-index).

**Results and Discussion**

*Testing the Research Hypotheses*

The first research hypothesis of this research paper concerns ‘the inter-rater reliability of holistic method is higher than the inter-rater reliability of PIE method’. To accept or reject this hypothesis, this research paper makes use of Intraclass Correlation Coefficient (inter-rater reliability) to examine which assessment method is higher.

**Table 4. Holistic Method**

<table>
<thead>
<tr>
<th></th>
<th>Intraclass Correlation</th>
<th>95% Confidence Interval</th>
<th>F Test with True Value 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Value</td>
<td>df1</td>
<td>df2</td>
</tr>
<tr>
<td>Single Measures</td>
<td>.363</td>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>Average Measures</td>
<td>.532</td>
<td>39</td>
<td>39</td>
</tr>
</tbody>
</table>

Two-way mixed effects model where people effects are random and measures effects are fixed.

- a. The estimator is the same, whether the interaction effect is present or not.
- b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.
- c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.
Table 5. PIE Method

<table>
<thead>
<tr>
<th>Intraclacel Correlation Coefficient</th>
<th>95% Confidence Interval</th>
<th>F Test with True Value 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Measures</td>
<td>.922&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.857 .958</td>
</tr>
<tr>
<td>Average Measures</td>
<td>.959&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.923 .979</td>
</tr>
</tbody>
</table>

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

According to tables (4) and (5), we fail to accept this hypothesis due to the fact that the raters of the holistic method did obtain a lower agreement (53.2%) on the rank orders of the subjects than the raters of the PIE method (95.9%). This seems that utilizing holistic method in evaluating translation results in a lower inter-rater reliability than PIE method. For instance, the translation daft of subject (28) received (18.33 and 18) from the trainer and professional raters in PIE method, while (16 and 18) from the holistic raters. This completely shows that raters in PIE method are more consistent compared to holistic method.

The second hypothesis concerns ‘the inter-rater reliability between the evaluators will be in between in the case of analytic method’. To accept or reject this hypothesis, we must look at how the raters did in terms of inter-rater reliability when they scored analytically. The following table shows the correlation coefficient (Spearman rho) between holistic and analytic raters. With reference to the first hypothesis, the raters in holistic method obtained a lower agreement compared to the raters of PIE method. The correlation coefficient ranges from 0.285 and 0.766. This seems to illustrate that the use of analytic method results in a higher inter-rater reliability compared to holistic method. Therefore, we can confirm that the interrater reliability is in-between in the case of analytic method (51.8%). But, the inter-rater reliability of analytic method is still far more than the reliability obtained through PIE method (Cronbach’s Alpha: 95.9%).

Table 6. Correlation Coefficient between Holistic and Analytic Raters

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Ho1</th>
<th>Ho2</th>
<th>An1</th>
<th>An2</th>
</tr>
</thead>
</table>

13
According to above table, we accept the hypothesis and the inter-rater reliability set of these three assessment methods are as follows:

**Holistic method<Analytic method<PIE method**

In the light the results, although spearman correlation coefficient were illustrated to be statistically significant for both holistic and analytic assessments, they did not echo a convincing inter-rater reliability in this assessment context. For instance, the highest correlation coefficient between holistic and analytic raters (r = 0.766) expatiates only (59\%) (r² = 58.67) of the variance between the raters. In other word, the prognostic value of the score of one rater for the outcome of another rater’s score is constrained. This indicates that the rank orders of the scores attributed to the same translation performance through holistic and analytic methods differ significantly (see table 6). For example, participant (7) got 17 and 13 from the holistic raters (trainer and professional), whereas he got 14.50 and 18 from the analytic raters. To take another example, participant (14) got 18 and 15 from the holistic raters and 14.50 and 16 from the analytic raters. This kind of scoring shows that notwithstanding the punctilious process of scoring of both raters (trainer and professional), the inter-rater reliability between the holistic and analytic raters are implausible. On balance, we can come to this conclusion that the subjects’ chance of prospering at a test would be highly associated with the rater in question on the basis of holistic and analytic assessments. If we look at the descriptive statistics among the three methods attributed to six raters, we figure out small difference in the obtained means and standard deviation.

**Table 7. Descriptive Statistics**

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>N Statistic</td>
</tr>
<tr>
<td>---</td>
</tr>
</tbody>
</table>

---

**. Correlation is significant at the 0.01 level (2-tailed).**
The third hypothesis concerns ‘due to the subjective results of both holistic and analytic methods, they will not be able to discriminate students’ performance. In case of PIE method, the discriminatory power is maximized through the preselected items in terms of ‘p-docimology’ and ’IDis-index’. This hypothesis illustrates the discriminatory power of PIE method. In this direction, PIE method is contrived to maximize the power discrimination, firstly through the preselection of items in the source text and then the calculation of docimological-p and d-index. To accept or reject this hypothesis, scores assessed by the PIE raters were compared to those of scores evaluated by holistic and analytic raters. PIE method comprises five stages. In the first phase, on the basis of evaluators’ experience, they distinguished all the eligible preselected items which may be the source of errors such as grammatical errors, additions, omissions, misinterpretations, literal translations, and lack of stylistics errors. They preselected twenty items according to their longstanding experience on evaluating translation. Secondly, they scored the translation drafts traditionally, yet on the basis of the preselected items in the source language. In the third phase, they calculated the docimological-p of the preselected items to see which item has a good p-value. According to Segers and Kockaert (2016), p-value should be higher than 0.20 and lower than 0.90 showing the item not too difficult and not too easy. Fourthly, in terms of docimological-p and the obtained scores, the evaluators identified the discriminatory power of the preselected items through extreme-group method or the twenty-seven percent rule to distinguish the upper scorers and the lowers one. Based on item discrimination (d-index), the evaluator selected any item which was 0.40 or higher as a very good item. In the fifth phase, on the basis of docimological justified items (good p-value and d-index) (Items 3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 16, and 17 were selected as the accepted items docimologically justified in this research), the evaluators re-calculated the scores (out of 20). According to the recalculation of the final scores, the lowest and highest scores belong to participant (1) and (29, 30, 33, 35, 38, 39, and 40), mean and standard deviation are (16.16/20) and (SD = 3.43) respectively. The consequence of this recalculation is the most critical for participant (1) going from 6.33 to 3.33/20. This is due to the fact that participant (1) has correctly rendered the preselected items(2, 7, 11, 12, 13, 15, 18, and 19). However, six of these items (2, 7, 12, 15, 18, and 19) (docimologically unjustified) are not accepted after the calculation of p-docimology and d-index of the preselected items. Docimological-p and d-index contribute to one study measured objectively and consequently give the objective picture of the readability of a translation test. In the light of the obtained results, we confirm this hypothesis in which PIE method maximizes the discriminatory power (r = 95.9%).
Figure 1 Scores Distribution by the Holistic Rater (Trainer)

Figure 2 Scores Distribution by the Holistic Rater (Professional)

Figure 3 Scores Distribution by the Analytic Rater (Trainer)
Figure 4 Scores Distribution by the Analytic Rater (Professional)

Figure 5 Scores Distribution by the PIE Rater (Trainer)

Figure 6 Scores Distribution by the PIE Rater (Professional)
According to figures above, scores distributions for the three methods are almost similar. This shows that holistic, analytic, and PIE methods perform well in recognizing lower-marked students from the higher ones in spite of different means. The means of these methods illustrate global effect and are the scores result calculated through the raters. If we scrutinize the distribution of individual scores carried out by the raters, the differences in scores are strongly relied upon the rater and evaluation method applied. To sum, any evaluation method which is evaluator-dependent such as holistic and analytic (except CDI) will illustrate global effect and the score results are depended on the evaluation of different raters. However, in the case of PIE method, the first two stages such as the preselection of items and determination of correct and incorrect answers are evaluator-dependent. Care must be taken to reduce the dependency through using different evaluators who decide by consensus about the preselection and the items. The other three stages in PIE such as the calculation of scores, determination of docimological-p and d-index, and the recalculation of scores are not evaluator-dependent; therefore they are not exposed to global effect.

In the upshot, to accept or reject the next three correlation hypotheses, this paper calculated the correlation strength of these methods by means of spearman rho.

*Table 8. Correlation Strength*

<table>
<thead>
<tr>
<th>Correlations</th>
<th>PIE</th>
<th>Holistic</th>
<th>Analytic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spearman's rho</strong></td>
<td><strong>Correlation Coefficient</strong></td>
<td>1.000</td>
<td>.620**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td><strong>Holistic</strong></td>
<td><strong>Correlation Coefficient</strong></td>
<td>.620**</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td><strong>Analytic</strong></td>
<td><strong>Correlation Coefficient</strong></td>
<td>.686**</td>
<td>.718**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.</td>
</tr>
<tr>
<td>N</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

According to the above table, on the basis of correlation strength (Rumsey, 2016) the correlations between the raters are as follow:

*Table 9. Correlation Strength*
Based the provided information from the intended table, we accept these three hypotheses. This illustrates that PIE, Analytic, and Holistic methods are consistent with each other. To prove this statement, the figures below indicate the scatterplots illustrating the strong uphill positive linear pattern in terms of strength and the direction of relationship. However, figure (7) shows that raters in holistic and analytic method are more concordant compared to holistic and PIE, and analytic and PIE. As far as the last three stages of PIE method are not evaluator-dependent, the results will not be affected by the influence of the raters. This is due to the fact that the calculation of p-docimology and d-index is not exposed to global effect. They are both subjective and highly dependent on the raters. Overall, they will not be able to discriminate the performance of the participants sufficiently.

**Figure 7 Scatterplot Correlation (Holistic/Analytic)**

**Figure 8 Scatterplot Correlation (Holistic/PIE)**
Conclusion

Implication

According to the obtained results, this research came to this conclusion that holistic and analytic assessments carried out by the raters lack discriminatory power and test reliability. With this in mind, analytic assessment seems to manage a higher level of inter-rater reliability compared to holistic assessment, since it follows some fundamental criteria. However, analytic method contains too much variability in detail. Although holistic and analytic methods have their own advantages, they fail to apply incessantly owing to the debatable constancy of criteria. Holistic and analytic assessments are based on subjective approaches and their reliability is under question. Therefore, the justification of scores is problematic. On the other hand, Preselected Items Evaluation (PIE) method inspects students’ performance realistically and consistently, since this method is based on the preselection of items. The significant merit of PIE method is to be applied in test standardization in a sense that it paves the way for a valid and reliable translation competence certification. The preselected items are examined on their docimological leverage on the basis of p-value and d-index. Tellingly, PIE method increases reliability with test validity and the correlation among the scores on the translation test. It also controls the docimological value of a translation test.

Limitations

Perhaps, the significant limitation of this study is the small number of participants at MA level. Also, the participants carried out the assignment by paper and pencil. Care must be taken to prepare the situation by allowing the participants perform their assignments by computer. Two important questions may arise: (1) “What is happened to the items that was not preselected or what does an evaluator do with the participant initiating the wrong answer.
for the non-preselected item?” and (2) “Is PIE method regarded as a time-consuming methods?” To provide the precise response for the first question, the evaluator observes the overall performance of the participant taking a test. If the participant is the only one among the total population initiating the incorrect response for the non-preselected items; therefore, the intended item can be discarded and not be included in the translation test. However, if that item has a good p-docimology and d-index; therefore, it must be included in the test. To answer the second query, we have to profess such truism that PIE method entails a large time speculation; however, to get the precise picture of the docimological quality of the test, every translator has to pay such price. Perhaps, the only solution is to automate PIE method by computer softwares (e.g. TranslationQ). In this vein, a software has to control and check the answers in the translation imported to it and prepares the list of correct-incorrect solutions of the preselected items.

Suggestions for Further Research

Some suggestions in order to augment the efficiency of PIE method are as follows:

1. What should be the ideal length of the source text? Number of the words?
2. What is the ideal number of the preselected items in the source text?
3. How is it possible to lessen the degree of evaluator-dependency?
4. Which norm values should be ideal for the p-docimology in PIE method? Some believe that 0.20-0.90 (Kockaert and Segers 2016), while other deem that 0.27-0.79 (Greenland et al. 2009)?
5. Is the Ebel norm beneficial in PIE method?
6. Which are the criteria for preselection of an item? Does evaluator expertise suffice?
7. Can the evaluation in PIE method be automated for every pair of language?

References:


*Alireza Akbari* (Corresponding Author)
PhD Researcher of Translation and Technology
Faculty of Translation and Technology
KU Leuven, Belgium
E-mail: Alireza.akbari@kuleuven.be

*Winibert Segers*
Assistant Professor of Translation and Technology
Faculty of Translation and Technology
KU Leuven, Belgium
E-mail: Winibert.segers@kuleuven.be