ONOMÁZEIN
Journal of linguistics, philology and translation

PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE
FACULTAD DE LETRAS

Accurate evaluation and consistent results: the case of the optimized version of the preselected items evaluation method

Alireza Akbari
University of Isfahan
Iran

Mohammadtaghi Shahnzari
University of Isfahan
Iran

Mahmoud Afrouz
University of Isfahan
Iran

ONOMÁZEIN 68 (June 2025): 65-103
DOI: 10.7764/onomazein.68.04
ISSN: 0718-5758

Alireza Akbari: Faculty of Foreign Languages, University of Isfahan, Iran. | E-mail: alireza.akbari@fgn.ui.ac.ir
Mohammadtaghi Shahnzari: Faculty of Foreign Languages, University of Isfahan, Iran. | E-mail: m.shahnazari@fgn.ui.ac.ir
Mahmoud Afrouz: Faculty of Foreign Languages, University of Isfahan, Iran. | E-mail: m.afrouz@fgn.ui.ac.ir

68
June
2025

Received: April, 2021
Accepted: September, 2021

Abstract

Replicable/reliable translation evaluation methods are the most rudimentary thrusts in translation testing. The present research paper was an attempt to contribute to the omnipresence of product quality evaluation. In doing so, this sketch compared two translation evaluation methods, namely the PIE (Preselected items evaluation) and the OPIE (Optimized version of preselected items evaluation). The PIE method as a perturbative testing technique is used to evaluate the product quality. Despite its ubiquity, it could not accurately measure product quality through its parameters: p-value and d-index. To address this critical issue, this paper introduced an optimized version of this method (OPIE) to precisely evaluate the product quality through the application of Feldt's pG-value and the corrected item-total correlation (rit-value). One-hundred translation participants and seven evaluators partook in this research. The evaluators were asked to score translation drafts through the PIE and OPIE methods. To measure up the degree of reliability, Pearson product-moment correlation and regression variable plots were applied. The results demonstrated that the OPIE method was more consistent/reliable based on docimologically acceptable items. Research limitations and implications were discussed.

Keywords: translation evaluation; product quality evaluation; PIE method; OPIE Method, p-value; d-index; rit value; reliability.

1. Introduction

Translation evaluation and assessment are the two vital sub-areas in Translation Studies (TS, hereafter), corroborated to be swarming with stumbling blocks. In the past decade, stumbling blocks appeared chiefly troublesome to translation agencies and translators. Both language service providers (LSPs) and freelance translators' objective was to hand over high-quality products to entice new customers/clients. The authority of TS has remained quality as a shielded distance for several years. During previous years, several research articles in translation evaluation and assessment have dealt with the issue of translation quality. Translation quality is one of the most discussed translation practice issues (Colina, 2009; Van Egdom and others, 2019). There is dissension over the issue of translation quality, but the way of assessing/evaluating quality, in general, and translation quality, in particular, is a hub of worldwide discussion among researchers/scholars (Larose, 1998; Williams, 2001). In the past decade, TS researchers have made their utmost efforts to compensate for this slanted circumstance and 'raise the profile of TS' (Van Egdom and others, 2019). This implies the demand for more empirical research on translation evaluation and assessment to hand over 'a sound methodological basis for the assessment practice' that would come up with the 'justification of test scores' (Eyckmans and Anckaert, 2017) in both professional and academic settings (Han, 2016; Akbari and Segers, 2017a).

The summon for quality became omnipresent in TS around the turn of the century; therefore, LSPs noticed moderately unpredicted researchers/scholars in their pursuit of quality. According to Van Egdom and others (2019: 27): "Although it need not work to the advantage of the theoretical branch(es) of TS, these scholars still had a vested interest in the assessment and evaluation of translation for the simple reason that most scholars are also teachers".

In recent years, much progress has been made in translation teaching and training on the scene to what is labeled professionalism (Nord, 1988; Delisle, 1990/1993/1998). Delisle made significant contributions to the notion of professionalism, placing much emphasis on 'learning outcomes' (Delisle, 1990). In this vein, learning outcomes were associated with the issue of competence, for which they soon altered the sphere of translator training/teaching (PACTE, 2016; Hurtado-Albir, 2017). To be labeled 'a successful translator', a translator must obtain competences pertinent to translation. Thus, there must be ways to evaluate/gauge translation competences. Translation competence has been defined as "the underlying set of knowledge and skills put into operation when translation a source text into a target text" (Eyckmans and Anckaert, 2017: 40). With the competence perspective in a complete turnaround, the bar has been elevated significantly for translator education and training.

First and foremost, the European's Master in Translation (EMT) framework has designed 'the leading reference standards for translator training and translation competence' (e.g., language and culture competences, translation competence, technology competence, personal and interpersonal competences, service provision competence) (EMT, 2017). According to the EMT framework, translation competence lies at the hub of the translation service provision competences. The EMT framework (2017: 4) is based on the surmise that

> Translation is a process designed to meet an individual, societal, or institutional need. It also recognizes that it is a multi-faceted profession that covers the many areas of competence and skills required to convey meaning (generally, but not exclusively, in a written medium) from one natural language to another, and the many different tasks performed by those who provide a translation service. The framework, therefore, considers that translator education and training at the Master's degree level should equip students not only with a deep understanding of the process involved but also with the ability to perform and provide a translation service in line with the highest professional and ethical standards.

Even though translation quality assessment and translation competence assessment are two issues that are not to be amalgamated, the former alludes to the assessment of the product. Both issues are often utilized equally in TS (Secară, 2005; Stejkal, 2006). This is to be elucidated by the equation that is instinctively postulated between a 'translation competence' and 'translation product'. Eyckmans and Anckaert (2017: 40) have pinpointed that

> In formative and summative examinations as well as in selection procedures, the translation product is traditionally seen as a reflection of the underlying competence that is, per definition, visible. In this respect, translation assessment greatly resembles the domain of (foreign) language assessment, where the many components that contribute to language proficiency also constitute a "black box" that can be accessed only through eliciting language performance.

The generalization issue finds its place in both examination and selection course of action since assessment is associated with performance. For instance, once a translation candidate is granted a certificate and opted for a translation position, the surmise is that the candidate not only has the potential to translate a particular text but also he/she is capable (competent) of translating other types of text. Nevertheless, making implications about students' competence in terms of one or more products is perilous. Such implications are conducted by measurement error, as is the case in all scientific realms. Therefore, it is significant to write down "the confidence that can be placed in these generalizations by looking into the reliability of the test scores" (Eyckmans and Anckaert, 2017: 41). But, queries in terms of reliability of assessment methods have endured largely unresolved since teaching/training and testing of translation competence have mostly "in the hands of practitioners rather than of translation scholars and researchers" (Akbari and Shahnazari, 2019: 2). Over the past few years, a number of published research article have spotted the need for "experimental and empirical evidence for the evaluation and quality of translation tests" (Anckaert and others, 2008; Akbari and Segers, 2017b) as well as a measure of translation competence (Waddington, 2004; Eyckmans and others, 2009; Han, 2016).

In this light, the purpose of the present paper is to compare two models of translation evaluation products, namely preselected items evaluation (PIE, hereafter) and optimized version of preselected items evaluation (OPIE, hereafter), so as to contribute to the prospect of a reliable assessment method. As are the case with PIE and OPIE methods, correct and incorrect solutions are itemized for each preselected item in the source text based on translation relevance and translation criterion-referenced assessment. The present research aims to measure the degree of reliability through the following hypothesis: *the quality of translation can be assessed/evaluated more reliably through the OPIE method than the PIE method*. Besides, it is mandatory to mention that this research paper earmarks a summative evaluation through which translators' competence level is evaluated (Stejkal, 2006). This is mainly due to the fact that summative evaluation is taken into account a hair-trigger state of affairs in which the necessity for a reliable assessment is regarded as more riveting. According to Eberly Center (2020), “the goal of summative evaluation is to *evaluate student learning* at the end of an instructional unit by comparing it against some standards and benchmarks. Summative evaluations are often *high-stakes*, which means that they have a high point value”.

2. State of the art

2.1. What is evaluation? A skimpy peek

Currently, evaluation covers the didactic system examinations (e.g., curricula) and extra-academic spheres (evaluation conducted by institutions and companies). Didactically speaking, the term “evaluation” is associated with measurement. Therefore, an evaluator is regarded as an arbiter and the person who is evaluated must accede to the authority of an evaluator who is sometimes either impartial or objective. Ergo, evaluation is an indispensable element of pedagogical/didactic enactments (PACTE, 2008) “no longer concerned only exclusively with examinations” (Martínez-Melis and Hurtado-Albir, 2001). Martínez-Melis and Hurtado-Albir (2001: 276) have noted that

> According to the latest research trends, it is the subjects (both the trainer and the trainee), rather than the objects of knowledge (the objectives and contents of the teaching program) that are the real protagonists of evaluation. Evaluation has become more creative; various evaluation models have been formulated, not only for the purpose of *instruction* (assimilation of the program content) but also for the purpose of *education* (development of individual aptitudes).

In a didactic context, terms such as “assessment” and “evaluation” must not be applied replaceably. Assessment and evaluation are considered concomitant terms in most cases; however, this research paper will not regard these two terms as synonyms and likewise makes a heedful separation between them. In terms of “timing, measurement focus, the relationship between administrator and recipient, findings and uses, ongoing modifiability of criteria and measures, standards of measurement, and the relation between objects of assessment and evaluation", Straight (2002) makes a clear distinction between assessment and evaluation. In timing, the assessment is associated with improving learning (formative), while the evaluation is concerned with gauging quality (summative). In the measurement focus, the assessment has to do with the learning process (process-oriented), while the evaluation has to do with the learning product (product-oriented). Considering the relationship between administrator and recipient, the assessment is considered "internally reflective for criteria", while the evaluation is "prescriptive for external benchmarks" (ibid.). Based on findings and uses, the assessment is diagnostic (focusing on improvement areas), while the evaluation is judgmental (focusing on overall scores and results). Based on the ongoing modifiability of criteria and measures, the assessment is resilient since it is exposed to different problems while the evaluation is fixed. In the case of standards of measurement, the assessment is "absolute", while the evaluation is "comparative" (e.g., comparing high competent students to low-competent ones). And finally, considering the relation between objects of assessment and evaluation, the assessment is "cooperative", while the evaluation is "competitive" (ibid.).

2.2. Translation evaluation methods

The most significant alteration having been spotted in the realm of Translation evaluation (TE, hereafter) over the past few years has been led the way through the initiation of different translation evaluation models. As the first TE model, the holistic method was believed a reasonable and precise evaluation method (Akbari and Shahnazari, 2019). The holistic evaluation method is considered an intuitive method because the overall quality of a translation is judged based on an evaluator's impression or appreciation (Mariana and others, 2015). For instance, an evaluator/rater considers a translation fair/acceptable, while another evaluator/rater considers the same translation unfair/unacceptable. Therefore, the application of this evaluation method must be carried out cautiously. Even though additional scrutinies are required, several research studies witness a high correlation between holistic scoring and in reasoned evaluation (Beeby, 2000). Beeby (2000: 185) has pointed out that "many experienced teachers rely on holistic and impressionistic methods. In fact, a recent study comparing intuitive, holistic evaluation and reasoned evaluation of the same translations showed a very high correlation between the two types of evaluation. However, these methods cannot provide the detailed information needed by trainers and trainees to further define competence and improving training and learning process and performance".

In the context above, the application of the holistic method is strenuous to sell. Even though this evaluation method should not be denigrated for having a frail mechanism against students' opprobrium, this method is revealed "on little more than an appeal to authority" (Anckaert and Eyckmans, 2007). Besides, as noted by Bahameed (2016), Akbari and Gholamzadeh (2017), and Akbari and Shahnazari (2019), this method is considered a lenient method because "it can give very little chance to see the individual differences among those many top students" (Bahameed, 2016: 146).

As the second evaluation method, the analytic method is associated with error analysis. In terms of validity and reliability, it seems to have outflanked the holistic evaluation method (Waddington, 2001). In the case of time management, the analytic method is more time-consuming compared to the holistic one; however, “analytic evaluation has the advantage that evaluators and (student) translators alike stand poised to gain a more tangible sense of what is right and/or wrong in a translation” (Kockaert and Segers, 2017; Van Egdom and others, 2019). The analytic method evaluates the overall quality of a translation based on text segments (e.g., words, phrases, clauses). Translation errors must be examined based on the evaluation grid criteria (a rubricated approach) (see table 1) (Eyckmans and others, 2013). According to Colina (2009: 240), a rubricated approach evaluates “components of quality separately, consequently reflecting a componential approach to quality; it is also considered functionalist and textual, given that evaluation is carried out to the function and the characteristics of the audience specified for the translated text”.

In the context above, the componential approach is foregrounded through the conformation of an evaluation grid model (a matrix including several error levels and error types) or a rubric. In doing so, an evaluator/grader must first identify translation errors and likewise s/he records pertinent information in the margin.

TABLE 1

Evaluation grid criteria (Eyckmans and others, 2013)

SENS (Meaning or sense)	Toute altération du sens dénotatif: informations erronées, non sens... J’inclus dans cette rubrique les oublis importants, c’est-à-dire faisant l’impasse sur une information d’ordre sémantique. (*Any deterioration of the denotative sense: erroneous information, nonsense, important omission.......*)	-1
CONTRESENS (Misinterpretation)	L’étudiant affirme le contraire de ce que dit le texte: information présentée de manière positive alors qu’elle est négative dans le texte, confusion entre l’auteur d’une action et celui qui la subit...... (*The student misinterprets what the source text says: information is presented in a positive light whereas it is negative in the source text, confusion between the person who acts and the one who undergoes the action.......*)	-2
VOCABULAIRE (Vocabulary)	Choix lexical inadapté, collocation inusitée (*Unsuited lexical choice, use of non-idiomatic collocations*)	-1
CALQUE (Calque)	Utilisation d’une structure littéralement copiée et inusitée en français (*Cases of a literal translation of structures, rendering the text into-French*)	-1
REGISTRE (Register)	Selon la nature du texte ou la nature d’un extrait (par exemple, un dialogue): traduction trop (in)formelle, trop recherchée, trop simpliste... (*Translation that is too (in)formal or simplistic and not corresponding to the nature of the text or extract*)	-0.5
STYLE (Style)	Lourdeurs, répétitions maladroites, assonances malheureuses... (*Awkward tone, repetition, unsuited assonances*)	-0.5

GRAMMAIRE (Grammar)	Erreurs grammaticales en français (par exemple, mauvais accord du participe passé, confusion masculin/féminin, accords fautifs... + mauvaise compréhension de la grammaire du texte original (par exemple, un passé rendu par un préesent...) 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*)	-0.5
OUBLIS (Omission)	Voir SENS (*See sense/meaning*)	-1
AJOUTS (Addition)	Ajout d'informations non contenues dans le texte (sont exclus de ce point les étouffements stylistiques. (*Addition of information that is absent from the source text (stylistic additions are excluded from this category)*)	-1
ORTHOGRAPHE (Spelling)	Erreurs orthographiques, pour autant qu'elles ne modifient pas le sens. (*Spelling errors, provided they do not modify the meaning of the text*)	-0.5
PONCTUATION (Punctuation)	Oubli ou utilisation fautive de la ponctuation. Attention: l'oubli, par exemple, d'une virgule induisant une compréhension différente du texte, est considéré comme une erreur de sens. (*Omission or faulty use of punctuation. Caution: the omission of a comma leading to an interpretation that is different from the source text is regarded as an error of meaning or sense*)	-0.5

Other examples of analytic evaluation method involve Canadian Language Quality Measurement System (SICAL) (with a focus on major and minor errors), the National Accreditation Authority for Translators and Interpreters in Australia (NAATI) (with a focus on the quality of language, error typology, and importance of errors), Council of Translators and Interpreters in Canada (CTIC) (comparable standard referenced model) (Williams, 2001), SAE J2450 (with a focus on the automotive service information translations and accommodating styles), LISA QA Model (based on error levels such as mistranslation, accuracy, terminology, language, style, etc.), ITR BlackJack which is now part of Capita plc (with a focus on linguistic accuracy, subject matter expertise, and technical function), QA Distiller (incorrect characters, missing figures, and form related properties of a text), and SnellSpell (spelling errors, compiling reports, importing dictionaries, and editing for TTX and SDL XLIFF files).

Although the analytic evaluation method may well be regarded as an optimal mode of action across universities, it is primarily believed that this evaluation method is not without defects. For instance, an evaluator only concentrates on the small text segment of the source language and s/he cannot have a thoroughgoing spectacle of the target text. This method is subjective since what might be considered a slight error for an evaluator might be a severe grammatical mistake for another one.

At first blush, the Preselected items evaluation (PIE) method as a summative assessment was devised by two KU Leuven professors, Hendrik J. Kockaert and Winibert Segers. In terms of timing and functionality, preselected items in the source text are limited. PIE method is considered a calibrated (examining the precision of a gauging instrument) and dichotomous (differentiation between correct and incorrect responses) method (Kockaert and Segers, 2017). PIE method preselects items based on two factors, namely *p-value* (difficulty of an item or item facility) ("the probability that examinees will get the item correct for educational/cognitive assessments or response in a keyed direction with psychological/survey assessments") (Thompson, 2017) and *d-index* (item discrimination) ("how well an item can differentiate between good candidates and less able ones") (Maxinity, 2020). Lei and Wu (2007: 527) have spotted that "results of an item analysis can help determine the minimum number of items needed for the desired level of score reliability or measurement accuracy". In this light, Kockaert and Segers (2017) have stated that an optimal p-value must be higher than 0.20 and lower than 0.90. With this in mind, 0.20 shows fewer participants responding to an item correctly (a difficult item), and 0.90 demonstrates the larger number of participants answering an item correctly (an easy item) (Eyckmans and Anckaert, 2017).

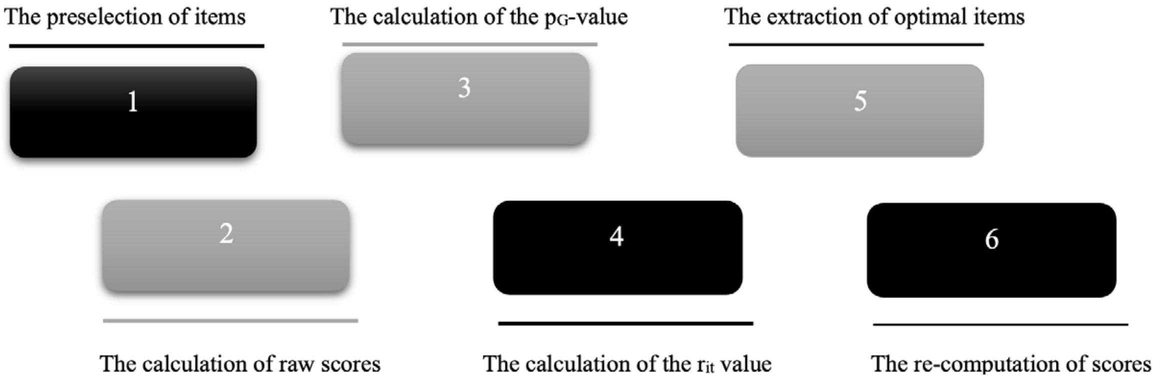
The extreme groups approach (EGA) calculates item discrimination (d-index) in the PIE method. According to Preacher and others (2005), "analysis of continuous variables sometimes proceeds by selecting individuals on the basis of extreme scores of a sample distribution and submitting only those extreme scores to further analysis. EGA is often used to achieve greater statistical power in subsequent hypothesis tests".

In the context above, the d-index can be measured through the EGA using the twenty-seven percent rule: twenty-seven percent of the top group of scorers and twenty-seven percent of the bottom group of scorers are analyzed. Wiersma and Jurs (1990: 145) have contended that "twenty-seven is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analyses". Statistically speaking, items having a d-index value of 0.40 are very good discriminators; items having a d-index value of 0.30-0.39 are considered good discriminators; items having a d-index value of 0.20-0.29 are fairly acceptable discriminators, and, lastly, items having a d-index value of 0.19 or less are considered weak/poor discriminators (Akbari and Segers, 2017a). With that in mind, items with an optimal p-value ($0.20 \leq p \leq 0.90$) and an optimal d-index ($d \geq 30$) are considered docimologically justified items and must be included within a test.

Translation scholars/researchers are now challenging to increase the reliability and validity of different translation evaluation methods. Therefore, proposing and formulating partially objective translation evaluation methods (not entirely objective methods) will improve translation quality assessment. In doing so, this paper introduces the optimized version of the PIE method, known as OPIE, which can be employed in academic (universities and institutions) and professional (translation service providers and companies with very developed expertise in translation evaluation and translation testing) settings. OPIE method consists of the following stages:

FIGURE 1

OPIE Stages


The preselection of items
1
The calculation of the pG-value
3
The extraction of optimal items
5
2
The calculation of raw scores
4
The calculation of the rit value
6
The re-computation of scores

Based on translation relevance, translation brief, domain criteria, and test specific criteria, the first stage is featured by preselecting items. Generally speaking, this stage is carried out before the test administration and examinees do not know which items are preselected. Evaluators can preselect items according to grammatical points, vocabularies, styles and so forth. The second stage is to calculate raw scores based on the holistic method. The scores provided in this stage are entirely based on evaluators' intuition. Evaluators are asked to score translation drafts based on Waddington's framework (2001) (see appendix 2). The third step is to calculate the degree of item difficulty using Feldt's pG-value ($0.55 \leq$ pG-value ≤ 0.67). The calculation of the pG-value is carried out the same as p-value of the PIE method; however, this parameter takes guessing factor into account. Acceptable items lie between 0.55 and .67. The fourth stage is to compute the degree of item discrimination using the r_{it} value (corrected item-total correlation). Acceptable/optimal items in this stage must be upper or equal to 0.30 ($r_{it} \geq 0.30$) (Anckaert and others, 2008). The fifth stage detects docimologically justified/optimal and calibrated items ($0.55 \leq$ *pG-value* ≤ 0.67 and $r_{it} \geq 0.30$). The last stage is to re-compute all scores based on the docimologically justified items.

3. Methodology

3.1. The purpose of the research paper

This research paper is an attempt to elucidate the full application of OPIE (a case study). It looks for testing/measuring the degree of reliability of the OPIE method compared to the PIE method.

3.2. Study condition and description of participants

Having signed letters of consent, one hundred translation students from the Bachelor of Arts at the University of Sheikhbahaei, Azad University, Shahreza Branch, and the University of Isfahan, Iran, partook in this research paper. Data were accumulated from different universities and at various learning/training phases to check both methods' reliability and validity. The distribution of participants are as follows:

TABLE 2
No of participants

NAME OF THE UNIVERSITY	BA, 3rd SEMESTER (6th)	BA, 4th SEMESTER (7th)	BA, 4th SEMESTER (8th)	TOTAL NO OF PARTICIPANTS
Sheikhbahaei University	10	8	12	30
Azad University	8	6	6	20
University of Isfahan	15	10	25	50
Total	33	24	43	100

The selected participants were all native Persian speakers (L1), and their ages range from 20 to 23. All participants were exposed to different translation text types such as journalistic, political, economic, and literary. Therefore, they had a clear insight into different translation courses. They were requested to translate a short economic excerpt (273 words) (see appendix 1) from English (L2) to Persian (L1) during the class hour under the supervision of the researcher and their instructors. Even though there were differentiations in their level of language proficiency, the quality standard was that it was by and large of a good command since the registration of their study programs called for passing prerequisite courses/units such as Advanced translation I and II, Arts and humanity translation, Simple text translation, and so forth. As mentioned, participants were given a short economic text and asked to translate it from English into plain Persian for an hour (60 minutes). This is an appropriate time for an average translator to translate a page containing 250 (in our case, 273 words) per hour (Sozonova, 2017). The participants were allowed to use whatever electronic dictionaries/resources as the authors tried to ensure that “heuristic competence was included in the translation performance” (Eyckmans and others, 2009). The subjects under study were all familiar with proper terminologies and structures of economic texts as they passed the relevant units associated with economic translation. The difficulty, length, and style of the given text were considered illustrative for the three universities' particulars. Eventually, the authors ordered three different reference translations of the short excerpt by three longstanding translation agencies (Basirat, Transnet, and Mandegar) and made them available to the evaluators. Therefore, the evaluators had access to a broad range of equivalents when evaluating the participants' translation drafts.

3.3. Procedure

Based on the quasi-experimental design, translation drafts were given to seven evaluators and asked them to score them (up to 20) based on the PIE and OPIE methods. The selection of evaluators was of great significance. Seven evaluators were selected based on their long-lived experience (nearly five years) in translation evaluation and testing. The evaluators were chosen from the University of Leuven (KU Leuven) and the Autonomous University of Barcelona. Three PIE evaluators were asked to score the translation drafts based on the PIE method principles (optimal p-value [0.20 ≤ items ≤ 90] and d-index [items ≥ 0.30]) (docimologically justified items). Also, four OPIE evaluators were requested to score the translation drafts using two parameters, namely (i) Feldt's pG-value (0.55 ≤ *pG-value* ≤ 0.67) and (ii) optimal r_{it} values (r_{it} ≥ 0.30) (docimologically justified items). For this research study, different types of statistical packages were used. To calculate the p-values and r_{it} values, the present research paper applied STATA (“a powerful statistical software that enables users to analyze, manage, and produce graphical visualization of data”) (Illinois Library, 2020) and SPSS (version 2017), respectively. Besides, to determine the degree of reliability of both methods, this research paper used the Spearman rank correlation coefficients and regression variable plots to check the relationship among the evaluators in terms of residuals/outliers.

4. OPIE Method: a case study

4.1. Stage 1: the preselection of items

Before evaluating, the evaluators were asked to preselected several items deemed appropriate. During this process, care was taken to linguistic and cultural differentiations as well as target text functions. Shortly after the preselection stage (commonly agreed items among evaluators), the possible Persian translations (reference translations) of the items were available to them. Twenty-seven items were commonly selected by the evaluators of KU Leuven and the Autonomous University of Barcelona (UAB). What was immediately apparent was that, once the evaluators disagreed with one another whether or not to put an item to the test, it was then preselected.

TABLE 3

Preselected items chosen by the evaluators of KU Leuven and UAB

NO OF ITEMS	PRESELECTED ITEMS	POSSIBLE PERSIAN TRANSLATIONS
1	developed financial systems	نظام مالی توسعه یافته/نظام مالی پیشرفته/ سیستم های مالی بهبود یافته
2	driving force	نیروی محرکه/نیروی پیش برنده/موتور رشد

3	Development	توسعه/پیشرفت
4	attenuating the costs	کاهش قیمت ها/ نزول قیمت/کاهش هزینه ها
5	such systems	این سیستم ها
6	financial intermediary	رابط مالی/ واسطه گری مالی/میانجیگری مالی
7	business opportunities	فرصت های سرمایه گذاری/ فرصت های تجاری
8	equipping savings	تجهیز منابع/ تجهیز پس اندازها
9	diversifying risks	متنوع سازی ریسک/ گوناگونی ریسک
10	services transactions	مبادلات خدمات/ تبادل خدمات/معامله خدمات
11	the ground for	زمینه/بستر/فضا
12	allotment of resources	تخصیص منابع/ تسهیم منابع/توزیع منابع
13	Accelerating	تسریع/رشد
14	capital accumulation	انباشت سرمایه/ اجمع آوری سرمایه/ انباشتگی سرمایه
15	macro levels	سطوح کلان/سطوح زیاد
16	notable point	مسئله حایز اهمیت/ مسئله مهم
17	distribution of wealth	توزیع ثروت/ توزیع درآمد
18	Inequality	نابراری/ناهمرتبگی/ناهمترازی
19	it is not clear	مشخص نیست/واضح نیست/معلوم نیست
20	leads to	منجر به/ منتهی به/باعث
21	Mainly	اساسا/عمدتا/علی الخصوص
22	Including	شامل
23	the point is that	موضوع این است که/ مسئله این است که
24	given that	با این فرض که/ بر این اساس/از آن جایی که
25	are responsible for	مسئول/ عهده دار/پاسخگوی
26	Apparently	ظاهرا/گویا/از قرار معلوم
27	However	با این وجود/ اما/ هرچند

4.2. Stage 2: the calculation of raw scores

One-hundred translation students were requested to translate a short economic text from English (L2) to plain Persian (L1). After administering the translation test, three professional evaluators (they belonged to the OPIE group) were asked to score the translation through Waddington's holistic framework (2001). The maintained scores were based on evaluators' anticipations/appreciations. The purpose of this stage is to show the huge difference between raw scores and OPIE scores done in stage 6. The scores are as follows:

TABLE 4

Raw scores

PAR	RAW SC.	PAR	RAW SC.	PAR	RAW SC.	PAR	RAW SC.
1	8.00	26	17.00	51	19.00	76	20.00
2	10.00	27	17.00	52	17.00	77	20.00
3	14.00	28	18.00	53	19.00	78	20.00
4	12.00	29	18.00	54	19.00	79	20.00
5	11.00	30	15.00	55	18.00	80	18.00
6	14.00	31	16.00	56	20.00	81	16.00
7	14.00	32	20.00	57	18.00	82	15.00
8	14.00	33	20.00	58	18.00	83	17.00
9	12.00	34	19.00	59	18.00	84	18.00
10	12.00	35	20.00	60	18.00	85	18.00
11	12.00	36	18.00	61	17.00	86	15.00
12	16.00	37	19.00	62	14.00	87	12.00
13	12.00	38	19.00	63	14.00	88	12.00
14	18.00	39	20.00	64	15.00	89	15.00
15	12.00	40	20.00	65	15.00	90	16.00
16	19.00	41	18.00	66	15.00	91	18.00
17	14.00	42	18.00	67	14.00	92	18.00
18	14.00	43	19.00	68	18.00	93	20.00
19	18.00	44	15.00	69	17.00	94	20.00
20	16.00	45	11.00	70	18.00	95	20.00
21	18.00	46	14.00	71	19.00	96	20.00
22	18.00	47	18.33	72	19.00	97	19.00
23	14.00	48	18.00	73	19.00	98	20.00
24	15.00	49	19.00	74	20.00	99	19.00
25	17.00	50	19.00	75	20.00	100	20.00

Students [1], [2], [3], [4]. [5], [13], [15], [45], [87], [88] got the lowest scores compared to all students. Based on the experts' [evaluators] remarks, these students applied the word-for-word (one-to-one correspondence) translation technique, which led to fuzzy meanings in the target translation. They could not adopt optimal translation techniques while translating. These students made critical translation errors (e.g., semantic and syntactic errors), which led their translation to be unclear and inaccurate.

4.3. Stage 3: the calculation of pG-value

Unlike other evaluation models such as PIE, Holistic, Analytic, which ignore the guessing role while scoring translation drafts, the focus of *pG-value* (G denoted guessing) is on guessing. Mind that an examinee favors his/her guessing ability [cf. competence] when translating. Measuring the degree of an item's difficulty considering the guessing parameter is of utmost importance.

The purpose of this stage is to determine the degree of an item's difficulty when considering guessing. Therefore, *pG-value* can be defined as the ratio of participants responding an item correctly to the whole population in the presence of guessing parameter.

$$P_G\text{-}value = \frac{\textbf{the ratio of participants responding an item correctly when guessing is present}}{\textbf{the total examination participants}}$$

In this direction, the total examination participants equals participants responding an item. Acceptable items are those situated between the level of 0.55 ≤ *pG-value* ≤ 0.67 where 0.55 shows an item's difficulty and 0.67 demonstrates an item's facility (easiness).

TABLE 5

The calculation of the pG-values

ITEMS	CORRECT ANSWERS	P_G-values
1	59/100	0.59
2	44/100	0.44
3	51/100	0.51
4	74/100	0.74
5	59/100	0.59
6	55/100	0.55
7	56/100	0.56
8	62/100	0.62
9	48/100	0.48

10	13/100	0.13
11	37/100	0.37
12	55/100	0.55
13	77/100	0.77
14	57/100	0.57
15	64/100	0.64
16	61/100	0.61
17	89/100	0.89
18	38/100	0.38
19	51/100	0.51
20	25/100	0.25
21	60/100	0.60
22	63/100	0.63
23	59/100	0.59
24	60/100	0.60
25	57/100	0.57
26	75/100	0.75
27	95/100	0.95

According to table 5, considering the guessing parameter, item 27 (p_{G27}-value = 95/100 = 0.95) is considered the least difficult item, while item 10 (p_{G10}-value = 13/100 = 0.13) is considered the most difficult item. Items such as 2, 3, 4, 9, 10, 11, 13, 17, 18, 20, 26, and 27 are excluded from the test because of inappropriate pG-values. Acceptable items (e.g., 1, 5, 24, etc.) in this stage are not considered an unlimited warranty for score computation since this research analyzes another factor called the r_{it} value (discriminatory power).

4.4. Stage 4: the calculation of r_{it} value

Corrected item-total correlation (known as r_{it} value) is used to check “the performance of the scenarios relative to the knowledge test as a whole” (Van Ham, 2018). According to Akbari (2019: 8), “this value [r_{it} 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. Moreover, the r_{it} value tells a researcher to what extent items are correctly answered by high-performing participants compared to low-performing participants”.

Good discriminatory items are used to discriminate the participants. To do so, items with a r_{it} value of 0.19 or less are considered poor discriminators and must then be excluded from the test. Items with a r_{it} value from 0.20 to 0.29 are regarded as fairly good items and are not used to discriminate between the high and low-performing participants. Items with a r_{it} value of 0.30 and above are good and very good items and must be included in the test (Eyckmans and Anckaert, 2017). Table 6 shows that items such as 1, 3, 10, 13, 20, and 26 (highlighted) must be excluded from the test due to the lower amount of the r_{it} value (below 0.30 %).

TABLE 6

The calculation of the r_{it} value

ITEMS	SCALE MEAN IF ITEM DELETED	SCALE VARIANCE IF ITEM DELETED	CORRECTED ITEM-TOTAL CORRELATION	CRONBACH'S ALPHA IF ITEM DELETED
1	14.3000	34.879	0.167	.847
2	14.3100	32.762	0.540	.835
3	14.3600	34.253	0.274	.844
4	14.3000	32.818	0.531	.835
5	14.2800	33.274	0.450	.838
6	14.2800	33.295	0.447	.838
7	14.3200	33.008	0.495	.836
8	14.2500	34.088	0.309	.842
9	14.2900	34.087	0.305	.843
10	14.2200	35.022	0.148	.847
11	14.3200	33.816	0.351	.841
12	14.3200	33.796	0.354	.841
13	14.3200	34.119	0.297	.843
14	14.3300	33.759	0.360	.841
15	14.3100	33.953	0.327	.842
16	14.3300	33.799	0.353	.841
17	14.3000	33.707	0.371	.840
18	14.3000	33.707	0.371	.840
19	14.2700	34.320	0.265	.844
20	14.2800	32.992	0.502	.836

21	14.2500	33.725	0.374	.840
22	14.2800	33.335	0.439	.838
23	14.2800	32.911	0.516	.835
24	14.2500	33.725	0.374	.840
25	14.2800	33.032	0.494	.836
26	14.2100	34.794	0.190	.846
27	14.3000	32.636	0.564	.834

4.5. Stage 5: the extraction of optimal items

Of the twenty-seven items preselected by the evaluators, only thirteen items were selected as docimologically justified items based on the pG-values and r_{it} values. Other items were excluded from the test due to their inappropriacy. The participants' scores will be computed through the selected items.

TABLE 7

Docimologically acceptable and calibrated items

ACCEPTABLE ITEMS	OPTIMAL p_G-VALUES	OPTIMAL R_{it} VALUES
5	0.59	0.450
6	0.55	0.447
7	0.56	0.495
8	0.62	0.309
12	0.55	0.354
14	0.57	0.360
15	0.64	0.327
16	0.61	0.353
21	0.60	0.374
22	0.63	0.439
23	0.59	0.516
24	0.60	0.374
25	0.57	0.494

4.6. Stage 6: the re-computation of scores

As mentioned, evaluators were firstly asked to score translation holistically. The holistic method is a common translation method across universities. Scores provided by this method are totally subjective and docimologically unjustified. In the last stage, we then compared raw scores and OPIE to check which method yielded consistent (replicable) and accurate scores. Scores run by the OPIE evaluators are as follows:

TABLE 8

Scores re-computation by four evaluators
OPIE scores by evaluator 1 (KU Leuven)

PAR	EVAL 1	PAR	EVAL 1	PAR	EVAL 1	PAR	EVAL 1
1	6.69	26	17.33	51	17.00	76	18.00
2	8.00	27	15.00	52	17.00	77	19.00
3	12.30	28	17.00	53	18.00	78	20.00
4	13.84	29	17.00	54	18.00	79	19.00
5	12.00	30	17.00	55	17.00	80	19.00
6	14.00	31	17.00	56	18.00	81	18.00
7	15.00	32	18.00	57	20.00	82	17.00
8	14.00	33	18.00	58	20.00	83	16.00
9	13.00	34	16.00	59	20.00	84	16.00
10	12.30	35	18.00	60	18.00	85	16.00
11	12.30	36	20.00	61	19.00	86	16.92
12	14.00	37	20.00	62	16.00	87	15.00
13	13.84	38	18.00	63	16.00	88	15.00
14	15.00	39	18.00	64	16.00	89	17.33
15	14.50	40	18.00	65	16.00	90	15.00
16	18.00	41	16.00	66	14.00	91	17.00
17	16.00	42	16.00	67	15.00	92	18.00
18	16.00	43	16.00	68	17.33	93	17.00
19	16.00	44	16.92	69	17.00	94	17.00
20	16.00	45	14.00	70	17.00	95	18.00

21	16.00	46	16.00	71	17.00	96	18.00
22	16.00	47	17.33	72	17.00	97	17.00
23	16.92	48	16.00	73	17.00	98	18.00
24	14.00	49	17.00	74	18.00	99	20.00
25	15.00	50	17.00	75	18.00	100	19.00

OPIE Scores by evaluator 2 (KU Leuven)

PAR	EVAL 2	PAR	EVAL 2	PAR	EVAL 2	PAR	EVAL 2
1	6.00	26	18.33	51	18.00	76	18.00
2	9.23	27	18.00	52	17.00	77	19.00
3	10.76	28	18.33	53	19.00	78	18.00
4	14.00	29	18.00	54	18.00	79	18.00
5	13.84	30	17.50	55	18.00	80	18.00
6	15.00	31	19.00	56	19.33	81	19.00
7	16.92	32	19.66	57	18.00	82	19.00
8	15.00	33	19.00	58	18.00	83	16.00
9	12.30	34	17.00	59	18.00	84	18.00
10	13.00	35	20.00	60	19.00	85	16.92
11	13.84	36	19.00	61	19.00	86	16.00
12	15.00	37	19.00	62	16.00	87	15.00
13	14.00	38	17.00	63	18.00	88	16.00
14	16.00	39	20.00	64	16.92	89	18.33
15	15.00	40	20.00	65	16.92	90	16.00
16	15.00	41	16.92	66	15.33	91	18.33
17	16.00	42	16.92	67	17.00	92	20.00
18	16.00	43	16.92	68	18.00	93	18.00
19	18.00	44	16.00	69	17.00	94	18.00
20	16.92	45	15.33	70	18.00	95	19.33
21	16.00	46	17.00	71	18.00	96	19.00
22	18.00	47	18.00	72	18.00	97	18.33

23	16.00	48	16.00	73	17.00	98	20.00
24	16.00	49	18.00	74	19.00	99	19.00
25	18.00	50	18.00	75	18.00	100	18.00

OPIE scores by evaluator 3 (UAB)

PAR	EVAL 3	PAR	EVAL 3	PAR	EVAL 3	PAR	EVAL 3
1	6.69	26	18.00	51	18.00	76	18.00
2	7.00	27	16.00	52	17.00	77	19.00
3	13.00	28	18.00	53	19.00	78	18.00
4	13.84	29	18.00	54	18.00	79	18.00
5	13.00	30	18.00	55	18.00	80	18.00
6	14.66	31	17.00	56	19.33	81	19.00
7	16.00	32	19.00	57	18.00	82	19.00
8	14.00	33	18.00	58	18.00	83	16.00
9	13.84	34	18.00	59	18.00	84	18.00
10	12.30	35	19.33	60	19.00	85	16.92
11	13.00	36	18.00	61	19.00	86	16.00
12	14.66	37	18.00	62	16.00	87	15.00
13	12.00	38	18.00	63	18.00	88	17.00
14	18.00	39	19.00	64	16.00	89	18.00
15	15.00	40	19.00	65	16.00	90	16.00
16	14.33	41	16.00	66	15.38	91	18.00
17	16.00	42	18.00	67	17.00	92	18.00
18	16.00	43	16.66	68	18.00	93	18.00
19	16.92	44	16.00	69	17.00	94	17.00
20	16.00	45	15.38	70	18.00	95	19.00
21	18.00	46	17.00	71	18.00	96	18.00
22	16.92	47	18.00	72	18.00	97	18.00
23	16.00	48	16.00	73	17.00	98	19.33
24	15.38	49	18.00	74	19.00	99	18.00
25	17.00	50	18.00	75	18.00	100	18.00

OPIE scores by evaluator 4 (UAB)

PAR	EVAL 4	PAR	EVAL 4	PAR	EVAL 4	PAR	EVAL 4
1	6.00	26	17.00	51	18.50	76	17.00
2	7.23	27	17.50	52	18.00	77	20.00
3	12.00	28	18.33	53	18.50	78	19.00
4	14.00	29	18.33	54	19.33	79	18.50
5	13.84	30	18.50	55	16.92	80	18.00
6	14.66	31	18.00	56	20.00	81	18.66
7	16.92	32	18.50	57	19.33	82	19.00
8	15.00	33	19.33	58	18.66	83	16.00
9	13.00	34	17.00	59	18.33	84	17.00
10	12.30	35	20.00	60	18.66	85	17.00
11	13.84	36	19.33	61	20.00	86	17.50
12	14.00	37	18.66	62	16.00	87	16.00
13	14.00	38	18.33	63	18.00	88	17.00
14	15.00	39	18.66	64	16.00	89	17.00
15	15.00	40	20.00	65	16.00	90	16.92
16	15.00	41	16.00	66	15.00	91	18.33
17	16.66	42	16.00	67	16.00	92	18.33
18	18.00	43	16.92	68	17.00	93	18.00
19	16.00	44	16.92	69	16.92	94	18.00
20	16.00	45	15.00	70	18.00	95	18.50
21	16.00	46	16.00	71	18.00	96	19.00
22	16.92	47	17.00	72	18.50	97	16.92
23	16.92	48	17.50	73	17.00	98	20.00
24	15.00	49	18.33	74	18.50	99	19.33
25	16.00	50	18.33	75	19.00	100	18.00

Table 8 illustrates the participants' scores based on each evaluator. In this light, the differences between the holistic scores (stage 2) and the recomputed scores (OPIE run) are at times perceptible. For example, there is a radical difference between the holistic scores and the OPIE ones of students [1], [3], [22], [50] coming crashing down after recomputing scores (8.00 vs. 6.69/6.00/6.69/6.00) (14 vs. 12.30/10.76/13/12) (18 vs. 16/16/16/18) and (19 vs. 17/18/18/18.33). Notwithstanding their translations' overall quality, these students could not render most of the docimologically calibrated items after calculating the pG-values and rit values.

5. Results (verification of the proposed hypothesis)

Hypothesis: *The quality of translation can be evaluated more reliably through the OPIE method than the PIE method.*

The purpose of the proposed hypothesis is to figure out the degree of reliability of the PIE and OPIE methods to check which of the mentioned methods is more consistent and generates similar results when employed over and over under the same circumstances. Consistency and stability of results in different conditions are indicators for reliability. To do this, this research applied Pearson product-moment correlation coefficient. The reason to choose this measure of strength is that this research paper applies continuous variables (evaluator's scores). According to Ursachi and others (2015: 681), “a general accepted rule is that α of 0.60-0.70 indicates an acceptable level of reliability, and 0.80 or greater a very good level. However, values higher than 0.95 are not necessarily good since they might be an indication of redundancy”.

TABLE 9

The degree of reliability of the PIE and OPIE methods

		Pearson correlation coefficient						
		PIE1	**PIE2**	**PIE3**	**OPIE1**	**OPIE2**	**OPIE3**	**OPIE4**
PIE1	Pearson correlation	1	.616**	.505**	.553**	.604**	.609**	.590**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	100	100	100	100	100	100	100
PIE2	Pearson correlation	.616**	1	.422**	.372**	.482**	.499**	.408**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000
	N	100	100	100	100	100	100	100

PIE3	Pearson correlation	.505**	.422**	1	.271**	.394**	.341**	.263**
	Sig. (2-tailed)	.000	.000		.006	.000	.001	.008
	N	100	100	100	100	100	100	100
OPIE1	Pearson correlation	.553**	.372**	.271**	1	.861**	.873**	.910**
	Sig. (2-tailed)	.000	.000	.006		.000	.000	.000
	N	100	100	100	100	100	100	100
OPIE2	Pearson correlation	.604**	.482**	.394**	.861**	1	.920**	.916**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000
	N	100	100	100	100	100	100	100
OPIE3	Pearson correlation	.609**	.499**	.341**	.873**	.920**	1	.925**
	Sig. (2-tailed)	.000	.000	.001	.000	.000		.000
	N	100	100	100	100	100	100	100
OPIE4	Pearson correlation	.590**	.408**	.263**	.910**	.916**	.925**	1
	Sig. (2-tailed)	.000	.000	.008	.000	.000	.000	
	N	100	100	100	100	100	100	100

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

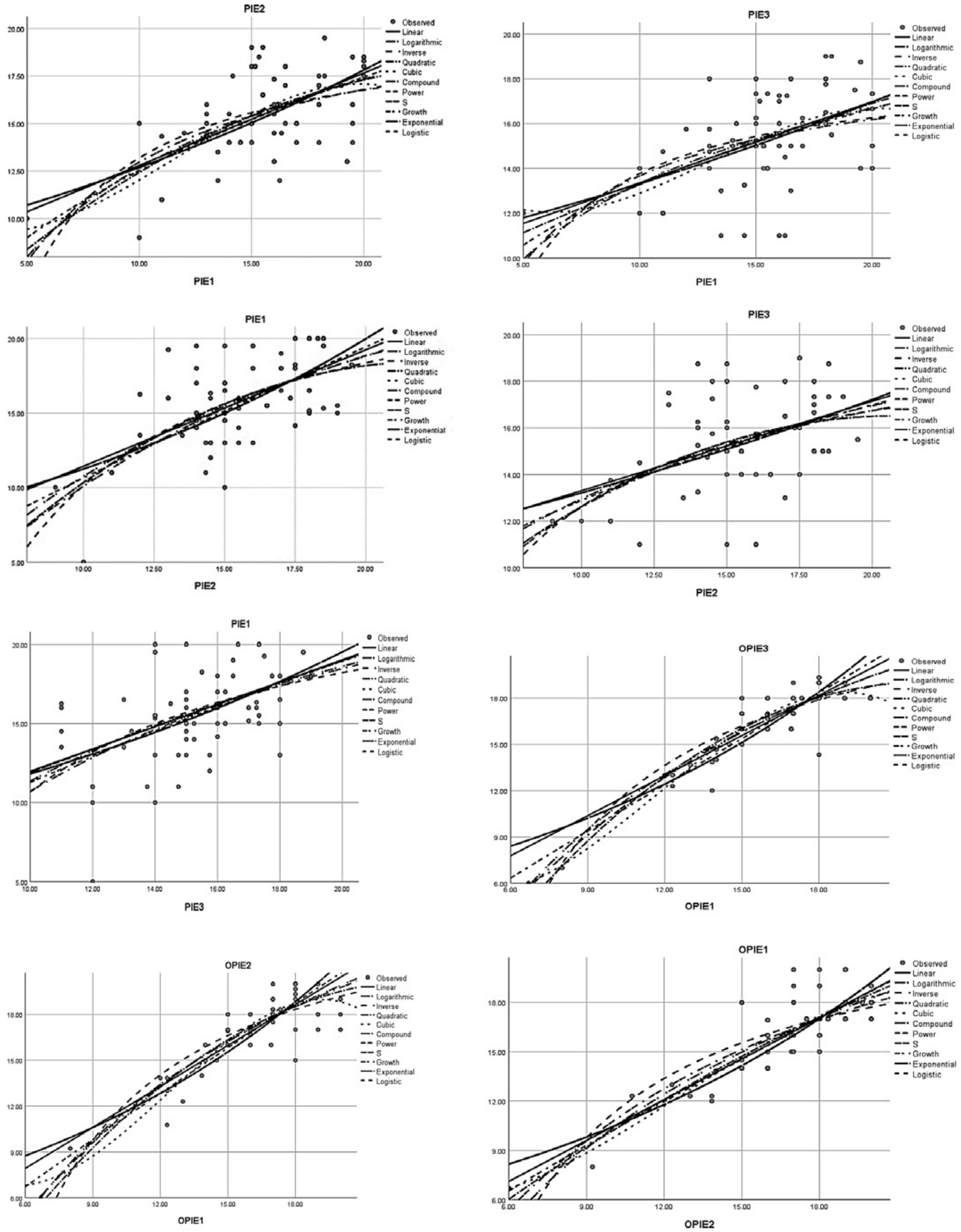
According to table 9, the interrater reliability results (also known as intercoder reliability) illustrated the OPIE evaluators/raters' precedence based on docimologically justified items. The results revealed that the OPIE method was more consistent/stable (e.g., 0.861, 0.873, 0.910, 0.920) compared to the PIE method (e.g., 0.616, 0.505, 0.422, 0.271).

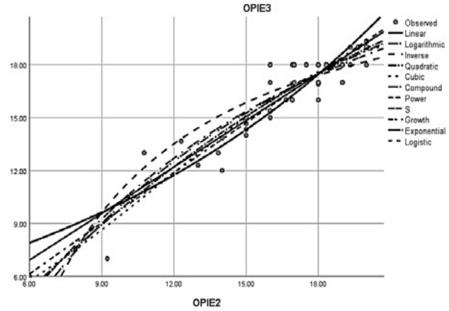
Regression variable plots (or added-variable plots) were applied to check the relationship between the response variable and one of the regression models' predictors. Gallup (2019: 1) has pointed out that “an added-variable plot is a scatterplot of the transformations of an independent variable (say, X1) and the dependent variable (Y) that nets out the influence of all the other independent variables. The fitted regression line through the origin between these transformed variables has the same slope as the coefficient on X1 in the full regression model, which includes all the independent variables”.

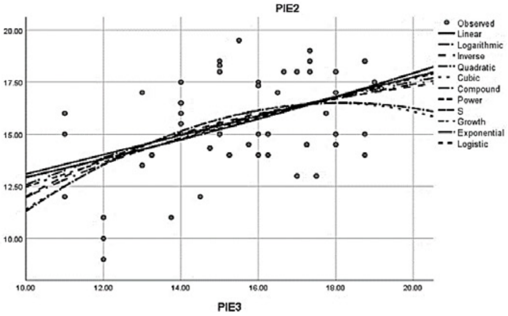
The following figure (figure 2) shows regression variable plots between the PIE and OPIE methods to illustrate the consistency among the raters:

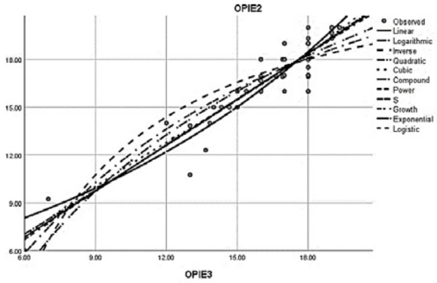
FIGURE 2

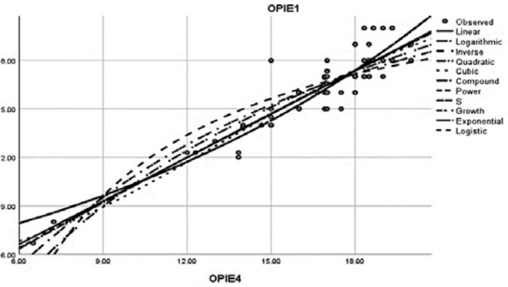
Added-variable plots of two methods

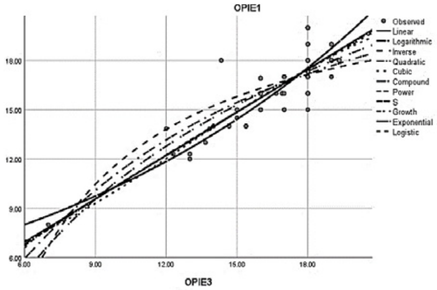


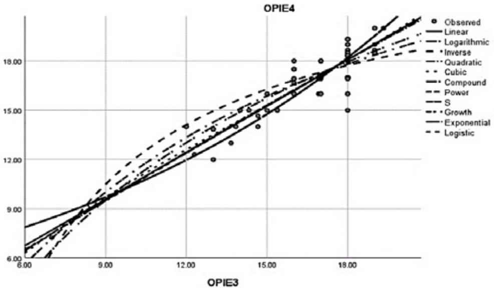

OPIE3
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE2

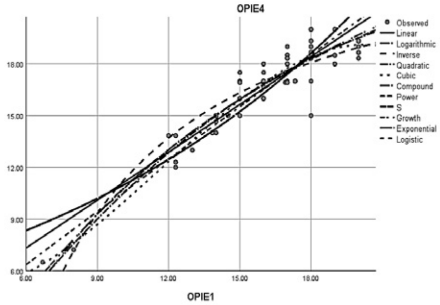

PIE2
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
10.00
12.50
15.00
17.50
20.00
10.00
12.00
14.00
16.00
18.00
20.00
PIE3

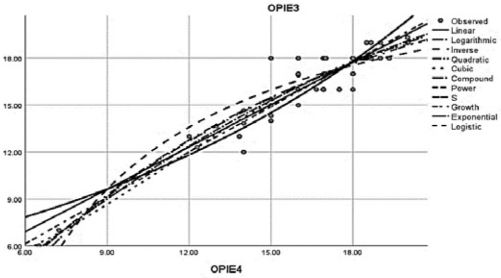

OPIE2
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE3

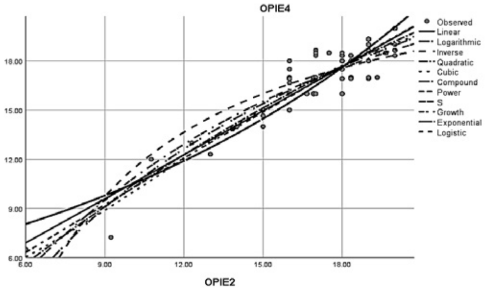

OPIE1
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE4

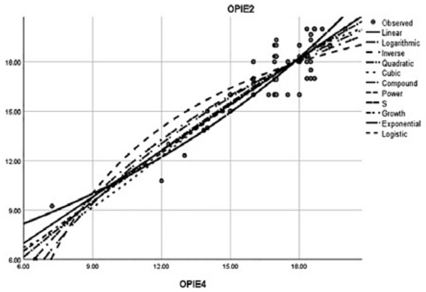

OPIE1
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE3


OPIE4
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE3


OPIE4
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE1


OPIE3
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE4


OPIE4
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE2


OPIE2
Observed
Linear
Logarithmic
Inverse
Quadratic
Cubic
Compound
Power
S
Growth
Exponential
Logistic
6.00
9.00
12.00
15.00
18.00
6.00
9.00
12.00
15.00
18.00
OPIE4

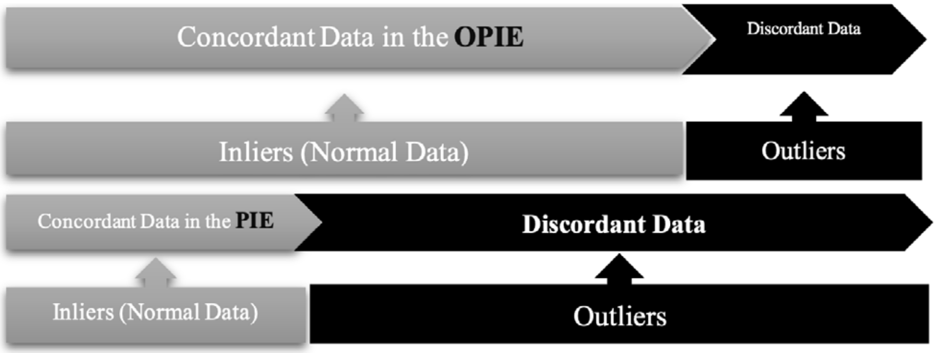
As can be observed, both methods represent several outliers. According to Aggarwal (2017), an outlier is a data point, which is remarkably different from the remaining data. Aggarwal (2017: 1) has also pinpointed that

> Outliers are also referred to as *abnormalities*, *discordants*, *deviants*, or *anomalies* in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves unusually, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities that impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights.

That being said, many outliers in a data point have negative impacts on the regression analysis and, likewise, they diminish the fit of the regression analysis. According to figure 2, OPIE raters/evaluators represented few outliers compared to the PIE method once scoring/evaluating the translation drafts. This illustrates that OPIE evaluators were more consistent and stable with one another; however, the PIE evaluators showed many outliers when scoring the translation draft. This demonstrated that the PIE method's application was not reliable enough to be used in both professional and academic settings, and it had negative and adverse impacts on the outcome of the test. Furthermore, the PIE method decreased the fit of the regression analysis.

FIGURE 3

Outlier spectrum between the OPIE and PIE



6. Discussion

6.1. P_G-values or P-values: which one holds a place?

Introductory measurement texts whip up instructors/university lecturers to design a test with a difficulty of 0.50. By difficulty, we mean the proportion of correct answers to an item made by students (i.e., p-value). There are no strong practical points to design items with various ranges of difficulty “even when the test is intended to rank students within a group” (Feldt, 1993). Even though the incorporation of items with different ranges/values of difficulty may be intriguing, the difficulty range of 0.50 and 0.60 will positively impact test reliability (Tinkelman, 1971; Feldt, 1993; Quaigrain and Kwamina Arhin, 2017). In this light, Sax (1989: 234-235) makes the following points:

(i) Items with the difficulty of 0.50 will augment observed score variance;

(ii) Items with the difficulty of 0.50 will yield the highest item discrimination;

(iii) The incorporation of items close to the difficulty of 0.50 will yield the highest test reliability.

To support this, Mehrens and Lehmann (1991) believed that items with difficulty of 0.50 are acceptable. They pointed out that the p-value “close to 0.95 or 0.05 fails to differentiate among students and hence cannot contribute significantly to the reliability of the test”. Besides, Tinkelman (1971) stated that a limited range of an item's difficulty situating from 0.50 to 0.65 is acceptable, where 0.50 refers to a difficult item and 0.65 refers to an effortless item. With that in mind, the optimal point of concentration of an item's difficulty is associated with another parameter called guessing (the probability of guessing a correct answer) (Draaijer and others, 2018), a case forgotten in the PIE method.

Both PIE and OPIE methods are associated with dichotomous items (0 shows an incorrect item and 1 shows a correct item/correct and incorrect answers). Dichotomization (a binary logic) in the PIE and OPIE methods yields two possibilities, namely (i) when the guessing parameter is not present and (ii) when the guessing parameter is present in a test. Feldt (1993: 38) has pointed out that “when there is no guessing, an instrument with item difficulties distributed symmetrically between 0.27 and 0.79 may be expected to have reliability only a few hundredths lower than a test with item difficulties concentrated at 0.50”.

In the context above, the PIE method founder adopted a different approach for an item difficulty. According to Kockaert and Segers (2017: 158), “p-values should be higher than 0.20 and lower than 0.90”. As it turns out, they did not substantiate whether or not the mentioned p-value range includes the guessing parameter. In doing so, it is mandatory to express one critical issue. In a test (in our case, a translation test), we cannot legitimately deny the guessing parameter. All students are speculated to be amply test-wise to guess while responding. In support of such a statement, Feldt (1993) has maintained that almost two-thirds of students guess at the answer. When guessing is present, three factors may be altered, namely, (i) inter-item correlation (“scores on one item are related to scores on all other items in a scale”) between items (Piedmont, 2014), (ii) covariance (the relationship between two random variables) between items, and (iii) the inclusive proportion of correct answers to each item. Considering this fundamental issue, the OPIE method focuses on measuring the degree of an item's difficulty concerning the guessing parameter through the p_G-value. For instance, participants [10], [13], [15], and [87] commonly scored 12 once the evaluators holistically scored the translation drafts (see table 4). However, when the OPIE method applied, their scores were changed to (Evaluator 1: 12.30, 13.84, 14.50, and 15) (Evaluator 2: 13, 14, 15, and 15) (Evaluator 3: 12.30, 12, 15, 15) and (Evaluator 4: 12.30, 14, 15, 16), respectively. This demonstrated that the guessing parameter affected the results.

In the context above, a question arises as to what extent the optimal range of difficulty will be if we consider a guessing parameter (p_G-value). To provide a clear-cut answer, several factors must be taken into account (1) the subject matter, (2) the intrinsic difficulty of an item, (3) Piedmont, 2014 guessing parameter, and (4) the resolution of a test that stirs up students to guess an answer. Considering these factors, the OPIE method took up Feldt's p_G-value ($0.55 \leq p_G\text{-value} \leq 0.67$) as the target value when guessing occurs in a translation test. Values outside the mentioned range of difficulty will not impact the test reliability.

6.2. A skirmish between OPIE and PIE's item discrimination: which one precisely discriminates high-performing students from low-performing students?

To discriminate high-competent translators from the low-competent ones, the PIE and OPIE methods adopt different approaches. The former utilizes the application of the extreme groups approach (EGA) (also known as extreme-group method) dating back to the pre-computer era, while the latter assigns corrected item-total correlation (r_{it} value).

As mentioned earlier, EGA is used to calculate the d-index. To calculate the EGA, Kockaert and Segers (2017) applied the 27% rule (i.e., the 27% upper and the 27% lower scorers). They believed that this amount would ‘maximize differences in a normal distribution’; however, this is not a case in point. D’Agostino and Cureton (1975) believed that each tail in a normal distribution must contain about 21% of the sample. D’Agostino and Cureton (1975: 47) have pointed out that

> In the traditional item analysis situation a test is administered to N subjects, the resulting scores are arranged in order of magnitude and the tests of these subjects whose scores fell in either the upper or lower tails of the distribution of test scores are further studied item by item. A question that arises here is how large the tails should be—that is, what is the appropriate *q* ($0 < q \leq 0.50$) such that the upper and lower tails to be selected for further study each contain *q* of the cases?

In doing so, they suppose that the distribution is normal and the sample size is large. The upper and lower tails' means of a sample are denoted by $\bar{X}_1$ *and* $\bar{X}_2$. Therefore, *q* must be determined so that the Critical Ratio (CR) is maximized;

$$CR = \frac{X1-X2}{SE\ (X1-X2)} \qquad (1)$$

In this direction, SE ($\bar{X}_1$-$\bar{X}_2$) is considered the standard error of $\bar{X}_1$-$\bar{X}_2$. Due to sampling fluctuation, equation (1) varies from sample to sample. Therefore, it is suitable to take anticipations and then maximize. So, we have:

$$E\ (CR) = \frac{2f\sigma/q}{SE\ (X1-X2)}\ (2)$$

F represents “the unit normal ordinate at the upper standardized baseline point”; *z* refers to the separation of a tail from the rest of the distribution, and *σ* demonstrates the standard deviation (SD) of the normal distribution. In this light, *z* and *f* are defined by

$$q = \int_z^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx\ (3)$$

And

$$f = \frac{e^{-z^2/2}}{\sqrt{2\pi}}\ (4)$$

Mosteller (1946) has pointed out that the upper and lower tails' observations are correlated for large samples. Furthermore, the SE of $\bar{X}_1$-$\bar{X}_2$ is incorrect and inappropriate and the q maximizing the amount of *CR* is not 0.27 percent. The correct and appropriate SE of $\bar{X}_1$-$\bar{X}_2$ is as follows:

$$SE\ (\bar{X}_1-\bar{X}_2) = {}_q\frac{\sigma}{\sqrt{N}}\sqrt{A}\ (5)$$

Where

$$A = 2q + 2zf + z^2 + (1-2q) - (2f + z\,(1-2q))^{\,2}\ (6)$$

The term to maximize is thus

$$E\ (CR) = \frac{2f}{\sqrt{A}}\sqrt{N}\ (7)$$

Or the last term is equivalent to

$$\frac{4f^2}{A}\ (8)$$

The last term can be calculated for *q* = 0.01(0.01)0.50.

D'Agostino and Cureton (1975) believe that *q* can be maximized around 0.21 and 0.22. Therefore,

q	***$4f^2/A$***
0.19	1.9293
0.20	1.9325
0.21	***1.9338***
Items are correlated in a normal distribution	
0.22	1.9336
0.23	1.9320
0.27	***1.9147***
Items are uncorrelated in a normal distribution	
0.50	1.7519

According to the normality, as the correlation escalates, the optimal tail size diminishes. This is mainly due to the fact that "if the concomitant variable and the test scores have correlation one, then the optimal tail size is around 0.215" (D'Agostino and Cureton, 1975: 49). In this light, the PIE method must reformulate its parameter (d-index) if done manually; so, the PIE results must be interpreted and analyzed with great care.

Although no reliability coefficients have been outlined in any PIE method accounts (Eyckmans and Anckaert 2017), the OPIE method calculates the item discrimination coefficient by applying the corrected item-total correlation (rit value). This value can be computed through various statistical software such as SPSS, STATA, and R. The rit value is applied to spell out "the association of the item with the total score on the other items" (Zijlmans and others, 2019). The OPIE method applies this value to figure out "the contribution of each item to instrument consistency as determined by the ability to discriminate between high and low-scoring individuals" (Wang and others, 2017). Likewise, the rit value contributes to the test's global reliability/replicability. In this light, the application of the rit value has exceptional merit compared to the PIE method's d-index in that the former analyzes every test-taker's score to calculate the discrimination coefficient; while the latter examines only 54 percent (i.e., the 27% upper and the 27% lower scorers) of test-taker's results by the EGA (Eyckmans and Anckaert, 2017: 45). The juxtaposition of the OPIE and PIE methods based on empirical data demonstrates the deficiencies of the latter. The PIE method does not discriminate the high-performing participants from the low-performing ones or does not discriminate between the test-takers' translation performances. Besides, the reliability of the PIE's test scores is under par (see table 9). Therefore, applying such a method in academic and professional settings must be carried out with great care.

7. Conclusion

7.1. Research limitations

Several limitations must be clarified when conducting this research paper, namely (i) small sample size, (ii) translation assignment, and (iii) timing. The present research paper's obtained results may not be sufficiently precise when juxtaposed to the large sample size (e.g., 400 participants). In further studies, larger samples are used (above 400 participants) to procure adequate and precise results. The second limitation is the translation assignment. The translation assignment (English to Persian) was conducted with a pen and paper. To save much time, care must be taken to design online platforms for participants to carry out their translation tasks by computers. Besides, in-depth knowledge of statistical packages such as SPSS and STATA is required for researchers/scholars to interpret and analyze results meaningfully. Last but not least, the application of the OPIE method is rather time-consuming. To clear up such a drawback, a computer platform is undoubtedly required to examine answers (removing subjective sways in the correction juncture) and provide a list of correct and incorrect responses/solutions (removing subjective sways in the selection juncture).

7.2. Research implications

Revisers, whether in professional and academic contexts, are cognizant of translation revision and scoring. Scoring and revising are considered repetitive and arduous tasks “where inconsistencies and subjectivity are hard to avoid” (Televic, 2020). Besides, an evaluator/reviser/rater cannot undertake to spot every mistake in all translation drafts, provide in-tune feedback and score/grade drafts congruously. To solve such problems, translationQ as a web and cloud-based platform is recommended. According to Televic (2020), “translationQ offers a new way of revising translations: let technology support you to increase consistency and objectivity, and provide a reliable trace of every error that was marked, score that was changed, and report that was generated”.

TranslationQ in action has several characteristics (in a nutshell):

(i) translationQ can be applied for every language pair and various text-types;

(ii) translation drafts can be imported in both word and SDL XLIFF files;

(iii) source texts can be aligned semi-automatically;

(iv) translation errors are tagged through customizable error categories;

(v) similar errors are detected automatically;

(vi) for every translation, general and specific feedback are provided;

(vii) analytics and insights are offered to revisers/evaluators;

(viii) translationQ is equipped with a user-friendly interface;

(ix) translationQ supports both left-to-right and right-to-left languages;

(x) for easy collaboration, revision memories are provided in TBX format, and

(xi) translationQ can be applied for high-stakes exams.

TranslationQ gives translation researchers/scholars and students the edge on the following issues, namely (i) “worry-free revision”: through new standards in translation revision and scoring, translationQ will act as your assistant to detect errors automatically, calculate scores, and generate reports; (ii) “intelligence under your control”: through the revision memory, a reviser can control the entire correction (accept or reject the detected errors) and feedback flow; (iii) “traceable objectivity”: the whole processes of translationQ are consistent, objective and traceable. An evaluator knows who rejects/accepts any errors and who overwrites the final score, and (iv) “productivity boost”: through automatic error detection, translationQ boosts evaluators' productivity “resulting in a reduction of the revision time” (Televic, 2020).

As long as the employment of the OPIE method is rather time-consuming, the automation of this method will add to the objectivity in translation testing and boost the scoring system efficiently. As mentioned, the task of the translationQ is to automatically identify predicted correct and incorrect solutions in the student versions. Therefore, the translationQ leads evaluators/raters/graders to get access to the vast range of correct and incorrect solutions/answers. Answers are carefully checked by evaluators and, in case of acceptance, the translationQ will automatically identify docimologically justified items (having optimal p_G-values and r_{it}-values). Then students' scores will pop up immediately and are sent to students.

After proposing the implementation of the OPIE method in the translationQ platforms, to boost the efficiency of the OPIE method, some questions may raise for which they need to be answered in further research, namely (i) what are ideal preselected items in a source text?, (ii) what criteria must be stipulated when preselecting an item?, (iii) what is an ideal length of the source text?, (iv) what are possible ways to reduce the control of evaluators and automate the preselection phase?, (v) what are the ideal number of evaluators and the level of their expertise when checking correct and incorrect solutions?, (vi) what would be the ideal number of justified items to reach a critical mass?, and (vii) would it be possible to calculate an item's difficulty (considering the guessing) through different approaches of item-response theory (e.g., one-two-three parameter logistic models)?

8. References

AGGARWAL, Charu C., 2017: *Outlier Analysis*, New York, USA: Springer International Publishing AG.

Akbari, Alireza, 2019: "Logistic calibrated items (LCI) method: does it solve subjectivity in translation evaluation and assessment?", *Revista de Lingüística y Lenguas Aplicadas* 14, 1-18, doi: 10.4995/rlyla.2019.11068.

Akbari, Alireza, and Monir Gholamzadeh, 2017: "Holistic Assessment: Effective or Lenient in Translation Evaluation?", *Skopos* 8, 51-67.

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

Akbari, Alireza, and Winibert Segers, 2017b: "Evaluation of Translation through the Proposal of Error Typology: An Explanatory Attempt", *Lebende Sprachen* 62 (2), 408-430, doi:10.1515/les-2017-0022.

Akbari, Alireza, and Mohammadtaghi Shahnazari, 2019: "Calibrated Parsing Items Evaluation: a step towards objectifying the translation assessment", *Language Testing in Asia* 9 (1), 1-27, doi: 10.1186/s40468-019-0083-x.

Anckaert, Philippe, and June Eyckmans, 2007: "IJken en ijken is twee. Naar een normgerelateerde ijkpuntenmethode om vertaalvaardigheid te evalueren" in C. Van de Poel and Winibert Segers (eds.): *Vertalingen objectief evalueren: Matrices en Ijkpunten*, Belgium: Acco.

Anckaert, Philippe, June Eyckmans and Winibert Segers, 2008: "Pour Une Évaluation Normative De La Compétence De Traduction", *ITL - International Journal of Applied Linguistics* 155 (1), 53-76, doi: 10.2143/ITL.155.0.2032361.

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

Beeby, Alison, 2000: "Evaluating the Development of Translation Competence" in Christina Schäffner and Beverly Adab (eds.): *Developing translation competence*, Amsterdam, NL: John Benjamins, 185-198.

Colina, Sonia, 2009: "Further evidence for a functionalist approach to translation quality evaluation", *Target: International Journal on Translation Studies* 21 (2), 235-264, doi: 10.1075/target.21.2.02col.

D'Agostino, Ralph B., and Edward E. Cureton, 1975: "The 27 Percent Rule Revisited", *Educational and Psychological Measurement* 35, 47-50.

Delisle, Jean, 1990: *L'enseignement de l'interpretation et de la traduction: De la theorie a la pedagogie*, Ottawa, Canada: University of Ottawa.

Delisle, Jean, 1993: *La traduction raisonnée: Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*, Ottawa, Canada: University of Ottawa.

Delisle, Jean, 1998: "Définition, rédaction et utilité des objectifs d'apprentissage en enseignement de la traduction" in Isabel García and Joan Manuel Verdegal (eds.): *Los estudios de traducción: un reto didáctico (Estudis sobre la traducció) (Spanish Edition)*, Castellón, Spain: Universitat Jaume I, 13-44.

Draaijer, Silvester, Sally Jordan and Helen Ogden, 2018: *Calculating the random guess score of multiple-response and matching test items* [https://eprints.soton.ac.uk/425109/1/Draaijer2018.pdf].

Eberly Center, 2020: *What is the difference between formative and summative assessment?* [https://www.cmu.edu/teaching/assessment/basics/formative-summative.html#:~:text=The%20goal%20of%20summative%20assessment,a%20midterm%20exam].

Eyckmans, June, and Philippe Anckaert, 2017: "Item-based assessment of translation competence: Chimera of objectivity versus prospect of reliable measurement", *Linguistica Antverpiensia* 16, 40-56.

Eyckmans, June, Philippe Anckaert and Winibert Segers, 2013: "Assessing translation competence", *Actualizaciones en Comunicación Social, Centro de Lingüística Aplicada, Santiago de Cuba* (2), 513-515.

Eyckmans, June, Philippe Anckaert and Winibert Segers, 2009: "The perks of norm-referenced translation evaluation" in Claudia V. Angelelli and Holly E. Jacobson (eds.): *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice*, Amsterdam/Philadelphia: John Benjamins, 73-93.

Feldt, Leonard S., 1993: "The Relationship Between the Distribution of Item Difficulties and Test Reliability", *Applied Measurement in Education* 6 (1), 37-48, doi: 10.1207/s15324818ame0601_3.

Gallup, J. L., 2019: *Added-variable plots with confidence intervals* [https://fmwww.bc.edu/repec/bocode/x/xtavplot_sj.pdf].

Han, Chao, 2016: "Reporting practices of rater reliability in interpreting research: A mixed-methods review of 14 journals (2004-2014)", *Journal of Research Design and Statistics in Linguistics and Communication Science* 3 (1), 49-75.

Hurtado-Albir, Amparo, 2017: *Researching translation competence by PACTE group*, Amsterdam, NL: John Benjamins.

Illinois Library, 2020: *Stata: Data Analysis and Statistical Software: Getting Started with Stata* [https://guides.library.illinois.edu/STATA].

Kockaert, Hendrik J., and Winibert Segers, 2017: "Evaluation of legal translations: PIE method (Preselected Items Evaluation)", *Journal of Specialised Translation* 27, 148-163.

Larose, Robert, 1998: "Méthodologie de l'évaluation des traductions", *Meta* 43 (2), 163-186, doi: https://doi.org/10.7202/003410ar.

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

Mariana, Valerie , Troy Cox and Alan Melby, 2015: "The Multidimensional Quality Metrics (MQM) framework: A new framework for translation quality assessment", *Journal of Specialized Translation* (23), 137-161.

Martínez-Melis, Nicole, and Amparo Hurtado-Albir, 2001: "Assessment In Translation Studies: Research Needs", *Meta* 46 (2), 272-287, doi: 10.7202/003624ar.

Maxinity, 2020: *Measuring Item Reliability Part 1 – Item Discrimination Index* [https://www.maxinity.co.uk/blog/item-discrimination-index].

Mehrens, William, and Irvin. J. Lehmann, 1991: *Measurement and evaluation in education and psychology*, New York, USA: Holt, Rinehart and Winston.

Mosteller, Frederick, 1964: "On some useful 'inefficient' statistics", *The Annals of Mathematical Statistics* 17, 377-408.

Nord, Christiane, 1988: *Textanalyse und Übersetzen: Theoretische Grundlagen, Methode und didaktische Anwendung einer übersetzungsrelevanten Textanalyse*, Heidelberg, Germany: Groos.

PACTE, 2008: "First results of a translation competence experiment: 'Knowledge of translation' and 'efficacy of the translation process" in John Kearns (ed.): *Translator and interpreter training: Issues, methods and debates*, London, UK: Continuum, 104-126.

PACTE, 2016: "Competence levels in translation: Working towards a European framework", *Paper presented at the 8th EST Congress*. Aarhus.

Piedmont, Ralph L., 2014: "Inter-item Correlations" in Alex C. Michalos (ed.): *Encyclopedia of Quality of Life and Well-Being Research*, Dordrecht: Springer Netherlands, 3303-3304.

Preacher, K. J., D. D. Rucker, R. C. MacCallum and W. A. Nicewander, 2005: "Use of the extreme groups approach: a critical reexamination and new recommendations", *Psychol Methods* 10 (2), 178-92, doi: 10.1037/1082-989x.10.2.178.

Quaigrain, Kennedy, and Ato Kwamina Arhin, 2017: "Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation", *Cogent Education* 4 (1), 1301013, doi: 10.1080/2331186X.2017.1301013.

Sax, Gilbert, 1989: *Principles of educational and psychological measurement and evaluation*, Belmont, CA: Wadsworth.

Secară, Alina, 2005: *Translation evaluation: A state of the art survey. Proceedings of the eCoLoRe/MeLLANGE Workshop Leeds*, Manchester, UK: St. Jerome.

Sozonova, Irina, 2017: *Translation turnarounds – How long does translation take?* [https://www.icanlocalize.com/site/2017/07/translation-turnarounds-how-long-does-translation-take/#:~:text=An%20average%20translator%20translates%20a,of%202000%20words%20per%20day].

Stejkal, Jiri, 2006: "Quality assessment in translation", *MultiLingual* 80 (17), 41-44.

Straight, H. S, 2002: *The Difference between Assessment and Evaluation* [https://www.binghamton.edu/academics/provost/assessment-and-analytics/assessment.html].

Televic, 2020: *TranslationQ: The game-changer in translation training and revision* [https://www.televic-education.com/en/translationq].

Thompson, Nathan, 2017: *What is classical item difficulty (P value)?* [https://assess.com/classical-item-difficulty-p-value/#:~:text=One%20of%20the%20core%20concepts,(more%20on%20that%20later)].

Tinkelman, S. N., 1971: "Planning the objective test" in . L. Thorndike (ed.): *Educational measurement*, Washington, DC: American Council on Education, 46-80.

Ursachi, George, Ioana Alexandra Horodnic and Adriana Zait, 2015: "How Reliable are Measurement Scales? External Factors with Indirect Influence on Reliability Estimators", *Procedia Economics and Finance* 20, 679-686, doi: https://doi.org/10.1016/S2212-5671(15)00123-9.

Van Egdom, Gys-Walt, Heidi Verplaetse, Iris Schrijver, Hendrik J. Kockaert, Winibert Segers, Jasper Pauwels, Bert Wylin and Henri Bloemen, 2019: "How to Put the Translation Test to the Test? On Preselected Items Evaluation and Perturbation" in Elsa Huertas-Barros, Sonia Vandepitte and Emilia Iglesias-Fernández (eds.): *Quality Assurance and Assessment Practices in Translation and Interpreting*, Hershey PA, USA: IGI Global, 26-56.

Van Ham, N. A. F., 2018: *Do employees in organizations still believe that conflict is always detrimental for team performance?*, Tilburg University.

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

Waddington, Christopher, 2004: "Should student translations be assessed holistically or through error analysis?", 49 (1), 28-35, doi: 10.1515/LES.2004.28.

Wang, Michael, Katharine Batt, Craig Kessler, Anne Neff, Neeraj N. Iyer, David L. Cooper and Christine L. Kempton, 2017: "Internal consistency and item-total correlation of patient-reported outcome instruments and hemophilia joint health score v2.1 in US adult people with hemophilia: results from the Pain, Functional Impairment, and Quality of life (P-FiQ) study", *Patient preference and adherence* 11, 1831-1839, doi: 10.2147/PPA.S141391.

Wiersma, William, and G. Stephen Jurs, 1990: *Educational measurement and testing*, Boston, USA: Allyn & Bacon.

Williams, Malcolm, 2001: "The Application of Argumentation Theory to Translation Quality Assessment", *Meta* 46, 326-344, doi: 10.7202/004605ar.

Zijlmans, Eva A. O., Jesper Tijmstra, L. Andries van der Ark and Klaas Sijtsma, 2019: "Item-Score Reliability as a Selection Tool in Test Construction", *Frontiers in Psychology* 9 (2298), doi: 10.3389/fpsyg.2018.02298.

Appendices

Appendix 1

According to economists like Schumpeter (1934) and Hicks (1985) developed financial systems are the driving force of economic development and growth at national level. Through attenuating the costs of supervision, transactions, and information, such systems play a key role in improvement of financial intermediary. By finding and funding business opportunities, equipping savings, covering and diversifying risks and facilitate goods and services transactions, developed financial systems prepare the ground for expansion of investment opportunities. Better allotment of resources, improvement of investment, and accelerating capital accumulation lead to higher economic growth and national revenue at macro levels. A notable point, however, is the effectiveness of financial development on distribution of wealth -i.e. income inequality and poverty. It is not clear if higher income and productivity are the only outcomes of financial development or it also leads to more just wealth distribution and attenuates income inequality and poverty? This is a critical issue mainly for the developing and even undeveloped countries including Muslim countries that claim to be the pioneers of expanding equality, social justice, and eradicating poverty. The point is that despite a desirable economic growth rate in some of these countries, which many factors are responsible for, inequality and poverty still are not eradiated in these countries. Given that financial development, apparently and like development of financial and credit tools, can provide the opportunity for poor and low-income classes to take part in economic activities; however, underdeveloped financial markets, failure of formation of domestic gross capital, absence of a political structure to support economic development, state's control on economy and small private sectors in Muslim countries add to the risk of using financial development.

Appendix 2

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 error.	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; through 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

Holistic Method of Assessment (Waddington, 2001)