Challenging the illusion of objectivity: an in-depth analysis of the preselected items evaluation (PIE) method in translation evaluation

Alireza Akbari and Mohammadtaghi Shahnazari
Department of English, Faculty of Foreign Languages, University of Isfahan, Isfahan, Iran

Abstract

Purpose – The primary objective of this research paper was to examine the objectivity of the preselected items evaluation (PIE) method, a prevalent translation scoring method deployed by international institutions such as UAntwerpen, UGent and the University of Granada.

Design/methodology/approach – This research critically analyzed the scientific and theoretical bottlenecks associated with the PIE method, specifically focusing on its parameters, namely the p-value and d-index, in adherence to established statistical protocols. Proposed remedies to mitigate the identified bottlenecks and augment the efficacy of the method were grounded in practicality.

Findings – The paper provided an extensive overview of the PIE method, which served as the foundation for the subsequent analysis and discussions. This research presented potential avenues for refinement and contributed to the current debate on objective translation assessment by addressing the theoretical and practical challenges associated with the PIE method.

Research limitations/implications – Translation researchers, practitioners and international institutions seeking to enhance the accuracy and reliability of translation evaluation should consider the implications of this research’s findings.

Originality/value – Although several publications focused on the role of the PIE method in translation evaluation, no study(ies) is available to critically analyze the scientific and theoretical bottlenecks of this translation evaluation method.

Keywords Reliability, Objectivity, p-value, d-index, Preselected items evaluation (PIE) method

Paper type Research paper

1. Introduction

Throughout the years, several research publications have been dedicated to exploring the topic of quality in translation assessment and evaluation. The quality issue within the field of translation studies (TS) is often regarded as a prominent subject of discussion in the realm of translation practice (House, 2013). The concept of quality and its evaluation in translation practice is debated. Translation academics and researchers have consistently emphasized the need for more empirical study on translation assessment to provide a solid methodological foundation for justifying translation test scores in academic and professional contexts (Han, 2016).

Translation competence and translation performance assessment are closely related. Translation competence, according to Postan (2020), includes a wide variety of skills and standards that go beyond a straightforward binary assessment of success or failure. It involves a detailed analysis of translation work, considering various dimensions and aspects of the translation process. Translation competence assessment and translation quality assessment (TQA) are the two terms that must not be confused (Eyckmans and Anckaert, 2017, p. 40). TQA is concerned with analyzing or rating final translation products (output
without considering the intricacies of the translation process) including translations produced by machine translation or computer-assisted translation tools. The goal of TQA is to assess the quality of the output by considering criteria including accuracy, coherence, fluency and adherence to the intended purpose and target audience (Han, 2020). As stated by Eyckmans and Anckaert (2017, p. 40),

In formative or 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, invisible. In this respect, translation assessment greatly resembles the domain (foreign) language assessment, where the many components contributing to language proficiency also constitute a “black box” that can only be accessed through eliciting language performance.

The assessment of translation competence, on the other hand, focuses on determining the underlying skills and abilities of the translator that support the translation process. It explores the translator’s ability with language, culture, subject matter, translation strategies and cognitive and problem-solving skills (Olalla-Soler, 2019). The assessment of translation competence aims to provide a comprehensive understanding of the translator’s overall capabilities and proficiency in producing high-quality translations (Galán-Mañas and Hurtado Albir, 2015).

The theme of generalization is included in examination and selection mechanisms due to the correlation between translation assessment and translation performance. For example, once a translation candidate obtains a certificate and opts for a position as a translator, the assumption is that the candidate has the potential to translate a particular text and is capable (competent) of translating other types of text. Nevertheless, making implications about students’ competence regarding one or more products is risky. As in all scientific fields, these implications are driven by measurement errors. Ergo, it is essential to jot down “the confidence that can be placed in these generalizations by looking at the reliability of the test results” (Galán-Mañas and Hurtado Albir, 2015). However, queries concerning the replicability of assessment approaches are left undecided. This is because the testing of translation competence assessment is essentially “in the hands of practitioners rather than translation scholars and researchers” (Akbari et al., forthcoming). Over the years, many published research articles have identified the obligation for experimental and empirical confirmation on the assessment and quality of translation tests (Akbari et al., forthcoming), as well as a gauging of translation competence (Waddington, 2004; Eyckmans et al., 2009; Han, 2016).

Given the above, the present research paper is an attempt to critically investigate one of the translation scoring methods called Preselected Items Evaluation (PIE hereafter), which is currently being tested and used through international institutions (e.g. UAntwerpen, UGent, the University of Granada). It is claimed that the PIE method objectively measures translation products (scores, in our case). The present article scrutinizes the PIE method’s scientific/potential bottlenecks through the following research questions:

- **RQ1.** What is the optimal degree of the \( p \)-value (item difficulty) in the PIE method?
- **RQ2.** What is the optimal degree of d-index (item discrimination) in the PIE method?
- **RQ3.** What alternatives can be applied to detect undocimologically justified (misfitting) items in the PIE method?
- **RQ4.** What alternatives can be applied to check the guessing parameter in the PIE method?

A short state-of-the-art concerning other translation evaluation (TE) methods is presented before poring over the research questions. This is followed by a general description of the PIE method, which will shape the point of discussion in this research article.
2. State of the Art

Within the domain of translation assessment and evaluation, there is a predominant emphasis on the detection and analysis of errors present in translated texts. Nevertheless, this approach requires a considerable allocation of human resources and involves substantial speculation, especially when confronted with significant quantities of translation drafts inside academic and professional contexts (Schmitt, 2005; Eyckmans et al., 2009). To tackle this particular difficulty, a range of evaluation methods have been used, such as the holistic and analytical methods (Conde Ruano, 2005). In recent years, the field of TE has seen significant changes due to the emergence of several methods to evaluate translations. The holistic method was the first TE model to emerge among the methods discussed, and it was initially regarded as a precise form of evaluation (Bahameed, 2016; Akbari et al., forthcoming).

The holistic method is mainly concerned with assessing the overall quality of a translation draft based on the evaluator’s subjective perception or appreciation (Mariana et al., 2015). Nevertheless, it is crucial to show prudence while using the holistic method, owing to its inherent subjectivity. For example, one evaluator may see a translation as suitable, while another may regard the same translation as unsuitable, depending on their subjective evaluation. Hence, it is essential to adopt the holistic method with meticulous deliberation. Although further study is required, several studies have shown a noteworthy association between the holistic scoring method and subjective evaluation (Beeby, 2000). This implies that while the holistic method is subjective in nature, it may provide valuable insights into the overall quality of a translation. Nevertheless, it is essential to recognize the want for more inquiry to ascertain the validity and reliability of this method in various circumstances. As believed by Beeby (2000, p. 185),

Many experienced teachers rely on holistic and impressionistic methods. 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 improve the training and learning process and performance.

In light of the aforementioned, it is challenging to convince individuals to embrace the holistic method. While this assessment method should not be disparaged for having a weak mechanism against students’ vilifications, it manifests itself “on little more than an appeal to authority” (Eyckmans and Anckaert, 2017). Additionally, as maintained by Bahameed (2016),

A disadvantage of this method is that it cannot easily distinguish the studious top respondents since their number may reach one-third of the whole translation class. That is to say, it can give top positions to many students, and this might give a negative impression that this method is too lenient to the extent that it can give very little chance to see the individual differences among those many top students.

The holistic method, as described by Garant (2009) and Mariana et al. (2015), evaluates the overall quality of a translation by assigning a score that is primarily determined by the evaluator’s overall impression and takes into account the final product in its entirety. In contrast, the analytic method analyzes individual segments of a translation and assesses it according to specific criteria and error categories (Waddington, 2001, p. 36). In the analytic method of TE, the quality of a translation is determined through an analysis of various text segments, including clauses, words and phrases. Using a rubric-based approach, translation errors are identified and analyzed in this fashion. As stated by Colina (2009, p. 240), a component-based approach to evaluating translation quality is reflected in the fact that a rubric-based approach permits the discrete assessment of quality components. The componental approach is prioritized in rubric-based evaluation by constructing an evaluation grid model or a rubric. These tools function as matrices comprising various
error levels and types, thereby furnishing evaluators with a methodical structure to analyze the translation. Generally, the grid model or rubric comprises predetermined categories and particular standards that correspond to the intended quality attributes and objectives of the translation task. A rater or evaluator identifies translation errors within the text segments prior to recording pertinent information in the margin or designated area when employing the analytic method. The information provided may comprise the error type and any supplementary remarks or recommendations for improvement. By methodically documenting these errors, evaluators can furnish the translator with precise feedback and direction for improving the quality of the translation.

It has been demonstrated that the analytic method exhibits superior validity and reliability compared to the holistic evaluation method (Waddington, 2001). Through the segmentation of the translation and the application of predetermined criteria, the analytic method facilitates a more methodical and comprehensive assessment, focusing on particular facets of the translation. This method furnishes evaluators with a systematic framework for assessing a multitude of linguistic and textual components, sense, style, register, etc. (Eyckmans et al., 2009). While the analytic method may require more time than the holistic method, it presents notable benefits. Using this opportunity, evaluators, translators, and students alike can acquire a more concrete comprehension of the elements that constitute right or wrong in a translation (Kockaert and Segers, 2017). The thorough review facilitated by the analytic method permits translators to receive targeted feedback and pinpoint domains that require enhancement. By adopting this method, one can develop a more profound comprehension of the translation process and acquire translation skills more efficiently.

Despite the fact that the analytic evaluation method may be regarded as the most effective method among institutions, it is generally accepted that this method is not devoid of errors. An instance of this would be when an evaluator concentrates on a minute portion of text in the source language and cannot examine the target text in its entirety. The grading process is inherently subjective, as what one rater may deem a minor grammatical error could be regarded as a major mistake by another (Akbari et al., forthcoming).

Scholars/Researchers have embraced new dispositions in previous years to freely assess “construct-irrelevant variables” (Eyckmans et al., 2009) ingrained in holistic and analytic evaluation methods. Some have chosen to scrap the search for an objective assessment/evaluation and examine themes such as intercoder and intra-coder reliability (Anckaert et al., 2008). Others have chosen to turn over “the methodology of measurement of education”, specifically the apprehension of Language Testing (LT), to the domain of TS and TQA by thriving norm-referenced translation tests (norm-referenced tests include the comparison of an examinee’s performance with that of other examinees) (Eyckmans and Anckaert, 2017; Eyckmans et al., 2009). The calibration of dichotomous items (CDI) method is the first attempt to determine translation competence based on the norm-referenced methodology (Anckaert et al., 2008). The CDI method comprises a norm-referenced assessment that aims to reveal discrepancies in “translation competence” among test-takers. The CDI is based on the bedrock that each text element putting up to the “measurement of differences in translation ability between test-takers” is given the status of an item (Eyckmans and Anckaert, 2017). The CDI method scores translation drafts as per a specific set of translated segments. The segments resulting from the CDI method are called “calibrated items” since the modus operandi for deciding them derives from LT and relates to how items are designed. The CDI is also a “dichotomous” method, which “precludes the weighing of mistakes” (Anckaert et al., 2008). This method bonds to a scoring system where translated segments are either permissible. This does not suggest only one suitable translation (solution) for the text segment; it simply implies agreeing on which options are permissible. Suppose there is contention about the admissibility of a translation alternative. In that case, the item should be
included in the procedure until the statistical evaluation of the pretest authorizes it to be excluded due to low discriminant power. By discriminating power, we mean which translated segments are rendered appropriately by examinees who flag up optimal translation competence (ability) and inappropriately by examinees, not competent translators.

3. Preselected item evaluation (PIE) method: an overview
The PIE method was initially implemented as a summative evaluation in the academic domain. The functionality and timeline of preselected items in the source text are limited. A calibrated method (one that examines the precision of a measuring instrument) and a dichotomous method (one that differentiates between correct and incorrect responses) are characteristics of this criterion-referenced evaluation approach (a criterion-referenced assessment assesses the performance of an examinee concerning a predetermined set of criteria). The PIE method is purported to provide a greater level of reliability in comparison to other evaluation methods, such as analytic and holistic methods, which rely on preconceived criteria and impressionistic-intuitive scoring (Kockaert and Segers, 2017, p. 160). Furthermore, there are assertions that this binary assessment method is suitable for all language pairs, including English-Dutch and English-French. The PIE method includes the following stages:

1. The preselection of some items in a source text as per evaluators’ skills;
2. The detection of correct/incorrect translations to the chosen items;
3. The calculation of item difficulty ($p$-value) (the proportion of examinees answering an item correctly) for each item;
4. The calculation of item discrimination ($d$-index) (candidates’ differentiations based on the items being measured) for each item and;
5. The recalculation of scores based on docimologically (acceptable) justified items’ difficulty values and items’ discrimination values (Colman et al., 2021).

As stated by Lei and Wu (2007, p. 527), the results of an item analysis ($p$-values) can assist in ascertaining the bare minimum number of items required to achieve the intended degree of score reliability or measurement accuracy. Kockaert and Segers (2017) concluded that to achieve an optimal $p$-value, it must be between 0.20 and 0.90. From this perspective, 0.20 indicates fewer participants accurately responding to a challenging item. In contrast, 0.90 indicates that more participants accurately responded to a simple item.

Conversely, the $d$-index illustrates how each item differentiates the highest and lowest scorers. It is expected that students who achieve a high score on the entire translation test will have correctly translated an item. With poor scores, on the other hand, it is anticipated that students have translated an item incorrectly. The calculation of the $d$-index is performed using the extreme groups approach (EGA) (Segers et al., 2018). As maintained by Preacher et al. (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.

The $d$-index can be calculated using the EGA by the 27% rule: set aside 27% of the groups with the lowest scores (lowest scorers) and highest scorers (top scorers). Wiersma and Jurs (1990) noted that “twenty-seven is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analyses.” Items with a $d$-index value of 0.40 or greater are regarded as excellent discriminators from a statistical
standpoint; those with a d-index value between 0.30 and 0.39 are deemed good discriminators; those with a d-index value between 0.20 and 0.29 are deemed moderately acceptable discriminators; and those with a d-index value of 0.19 or lower are regarded as weak or poor discriminators (Akbari et al., forthcoming; Colman et al., 2021). Based on this perspective, items that possess an optimal d-index (d-index ≥ 0.30) and p-value (20 ≤ p-value ≤ 90) are considered “docimologically justified items” (Akbari et al., forthcoming) and are required to be incorporated into the test.

4. Analyzing research questions

RQ1. What is the optimal degree of the p-value (item difficulty) in the PIE method?

Instructors are advised by introductory measurement texts to design a test with a difficulty level of 0.50. We refer to “the proportion of correct answers to an item provided by students (i.e. the p-value)” as “item difficulty” (Akbari et al., forthcoming). There are no compelling practical justifications for incorporating varying degrees of difficulty into items or tasks, even when the purpose of the assessment is to group students (Feldt, 1993). Although the inclusion of items with varying degrees of difficulty may be intriguing, research indicates that a difficulty range of 0.50–0.60 has a positive impact on test reliability (Feldt, 1993). As reported by Sax (1989, p. 234),

1. Items with a difficulty of 0.50 will augment observed score variance;
2. Items with a difficulty of 0.50 will yield the highest item discrimination;
3. The incorporation of items close to the difficulty of 0.50 will yield the highest test reliability.

According to Mehrens and Lehmann (1991), items with a difficulty level of 0.50 are deemed acceptable or suitable. They noted that a p-value “close to 0.95 or 0.05 fails to differentiate among students and therefore cannot significantly contribute to the reliability of the test.” Moreover, according to Tinkelman (1971), it was deemed appropriate for an item to have a restricted range of difficulty spanning from 0.50 to 0.65. In this context, a value of 0.50 signifies a challenging item, while 0.65 indicates a simple item. However, it is worth noting that the PIE method fails to account for the fact that the optimal concentration point of an item’s difficulty is dependent on an additional factor known as guessing (or the probability of correctly guessing an answer). The PIE method depends on dichotomous items where zero demonstrates an incorrect response, and one shows a correct answer. The dichotomy in the PIE method submits two probabilities: (1) when the guessing element is not present and (2) when the guessing element is present in a test. According to Feldt (1993, p. 38),

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 this regard, the PIE method incorporated a variety of approaches to the difficulty of an item. As stated by Kockaert and Segers (2017, p. 158), p-values should fall within the range of 0.90 to 0.20. Including the guessing element in the p-value range has not been demonstrated. It is not possible to validly eliminate the element of guessing in a translation test. It is presumed that all students are primarily attempting to guess when responding to an item. Feldt (1993) substantiated this assertion by stating that approximately two-thirds of students guess the correct answer. Three factors are susceptible to change in the presence of the guessing
element: (1) inter-item correlation, which states that “scores on one item are associated with scores on all other items in a scale”; (2) the covariance, which refers to the relationship between two random variables and (3) the inclusive proportion of correct responses for each item (Piedmont, 2014). If we contemplate a guessing element ($p_G$-value, where “$G$” represents speculating), an inquiry may emerge regarding the optimal difficulty range. Several factors must be considered to provide an accurate response: (1) the subject matter, (2) the intrinsic difficulty of a particular item, (3) the element of guessing and (4) the format of a test that encourages students to guess answers. When generating guesses for these factors, the PIE method must utilize the Feldt $p_G$-value ($0.55 \leq p_G$-value $\leq 0.67$) as the target value. The reliability of the test will not be impacted by values that fall outside the specified difficulty range.

Besides, applying the $p$-values in the PIE method bears several bottlenecks. They are as follows: (1) lack of differentiation: differentiating between items of similar difficulty levels may prove to be an ineffective task for the item difficulty index. Differentiating the relative levels of difficulty of multiple items becomes a challenging task when their $p$-values are similar. This lack of differentiation can limit the precision and discriminative power of the item difficulty index (Chauhan et al., 2023), (2) limited information for item improvement: merely depending on the item difficulty index does not furnish comprehensive insights into improving or revising an item. While it does indicate the level of difficulty or ease associated with an item, it does not reveal particular qualities or attributes of the item that could be amended to increase its quality or measurement properties (Tsaousis et al., 2018), (3) sensitivity to item format: The presentation or format of an item can influence the item difficulty index. Difficulty values can vary across item formats, even when evaluating the same underlying construct. The inability to compare and maintain consistency in item difficulty estimates across various assessment formats may result from this sensitivity to item format (Hontangas et al., 2015), (4) potential biases: numerous potential sources of bias could affect the item difficulty index, including cultural biases, language proficiency biases and item presentation biases. The accurate representation of the true difficulty level for all examinees may be compromised if an item exhibits bias towards a specific group or necessitates domain-specific or cultural knowledge (DeCarlo, 2020), and (5) inadequate for assessing item quality: Notwithstanding the significance of item difficulty as a determinant of quality, it fails to furnish an in-depth assessment of an item’s overall quality. While assessing the overall quality and efficacy of an item in precisely measuring the intended construct, additional factors, including item clarity, relevance and engagement, are equally vital (Donelan et al., 2016).

**RQ2. What is the optimal degree of d-index (item discrimination) in the PIE method?**

The PIE method employs the EGA to distinguish highly competent translators/students from those who are less competent. The d-index is computed utilizing EGA. To determine the EGA, Kockaert and Segers (2017) utilized the 27% formula, which stipulates that 27% of the strongest group (consisting of top scorers) and 27% of the weakest group (consisting of bottom scorers) be utilized. They believed this amount would “maximize differences in a normal distribution”; however, this is not the case. D’Agostino and Cureton (1975) postulated that in a normal distribution, each tail should encompass approximately 21% of the sample (Table 1). D’Agostino and Cureton (1975) have thought 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 fall 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 contains $q$ of the cases?
Assume that the sample size is large and the distribution is normal. \( \bar{X}_1 - \bar{X}_2 \) denote the upper and lower tails’ means of a sample. Therefore, \( q \) must be determined so that the critical ratio (CR) (a ratio that is related to a sample’s probability) is maximized;

\[
CR = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}
\]

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

\[
E(CR) = \frac{2q \sigma / q}{SE(X_1 - X_2)}
\]

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

\[
q = \int_{z}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx
\]

and

\[
f = \frac{e^{-z^2/2}}{\sqrt{2\pi}}
\]

Mosteller (1946) has noted that a correlation exists between observations of the upper and lower tails in large samples. Furthermore, the standard error of \( \bar{X}_1 - \bar{X}_2 \) is wrong and inappropriate, and the \( q \) that maximizes the CR amount is not 0.27%. The correct and reasonable SE of \( \bar{X}_1 - \bar{X}_2 \) is as follows:

\[
SE(\bar{X}_1 - \bar{X}_2) = q \frac{\sigma}{\sqrt{N}} \sqrt{A}
\]

where

\[
A = 2q + 2zf + z^2 + (1 - 2q) - (2f + z(1 - 2q))^2
\]

The term to maximize is thus

\[
Q = \frac{2\sigma}{SE(X_1 - X_2)}
\]

Items are correlated in a normal distribution

<table>
<thead>
<tr>
<th>Items are correlated in a normal distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q )</td>
</tr>
<tr>
<td>0.22</td>
</tr>
<tr>
<td>0.23</td>
</tr>
<tr>
<td>0.27</td>
</tr>
</tbody>
</table>

\( f \) shows “the unit normal ordinate at the upper standardized baseline point”;

\( z \) alludes to the separation of a tail from the rest of the distribution and \( \sigma \) shows the standard deviation (SD). In this light, \( z \) and \( f \) are defined by

\[
q = \int_{z}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx
\]

and

\[
f = \frac{e^{-z^2/2}}{\sqrt{2\pi}}
\]

Mosteller (1946) has noted that a correlation exists between observations of the upper and lower tails in large samples. Furthermore, the standard error of \( \bar{X}_1 - \bar{X}_2 \) is wrong and inappropriate, and the \( q \) that maximizes the CR amount is not 0.27%. The correct and reasonable SE of \( \bar{X}_1 - \bar{X}_2 \) is as follows:

\[
SE(\bar{X}_1 - \bar{X}_2) = q \frac{\sigma}{\sqrt{N}} \sqrt{A}
\]

where

\[
A = 2q + 2zf + z^2 + (1 - 2q) - (2f + z(1 - 2q))^2
\]

The term to maximize is thus

\[
Q = \frac{2\sigma}{SE(X_1 - X_2)}
\]

Items are uncorrelated in a normal distribution

<table>
<thead>
<tr>
<th>Items are uncorrelated in a normal distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0.50 )</td>
</tr>
</tbody>
</table>

Source(s): Table Courtesy of D’Agostino and Cureton (1975, p. 49)
\[ E(CR) = \frac{2f}{\sqrt{A}} \sqrt{N} \]  

(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) assert that \( q \) is maximized about 0.21 and 0.22. Hence, by the principle of normality, optimal tail size diminishes with increasing correlation. This is because the optimal tail size is approximately 0.215 “if the concomitant variable [1] and the test scores have a correlation of one” (D’Agostino and Cureton, 1975, p. 49). Given this consideration, manual implementation of the PIE method necessitates a reformulation of the \( d \)-index; thus, extreme caution is required when interpreting and analyzing the PIE results. Irrespective of the technique employed to compute the \( d \)-value, the PIE method yields discrimination indices from unconventionally small samples (Colman et al., 2021). Consequently, a cautious interpretation of the derived indices is required (Eyckmans and Anckaert, 2017). As expressed by Eyckmans and Anckaert (2017, p. 45),

The authors’ particular interpretation and application of the norm-referenced principle as about the performance of only a small group of tested candidates (a homogenous group of 12 master’s students in Segers and Kockaert (2016) and 19 master’s students in Kockaert and Segers (2014, 2017) has as a consequence that the generalizations that come with “true” norm-referenced testing, that is, on the basis of a representative sample of test-takers, become unavailable. Discrimination indices based on non-representative samples cannot be interpreted to represent real differences in translation competence.

To overcome this alarming barrier, the PIE method must compute the item discrimination coefficient by utilizing the corrected item-total correlation (rit-value). This value is used to express “the association of the item with the total score on the other items” (Zijlmans et al., 2019). The PIE method must utilize this value to ascertain “the contribution of each item to instrument consistency as determined by the ability to discriminate between high and low-scoring individuals” (Wang et al., 2017). Given this perspective, the utilization of the rit-value is markedly superior to the \( d \)-index of the PIE method, as the latter analyzes a mere 54% of test-takers’ scores (i.e. 27% of high scorers and 27% of low scorers) (Eyckmans and Anckaert, 2017, p. 45). As a result, the PIE method fails to distinguish between individuals who achieve the highest and lowest scores or evaluate the translation performance of the test takers.

Moreover, applying the \( d \)-index has several bottlenecks which should be taken into account. They are as follows: (1) sample dependency: the item discrimination index is influenced by the characteristics of the particular sample used in the analysis. Different samples may yield different discrimination values for the same item. This sample dependency can limit the generalizability of the index and raise concerns about its reliability across different populations or contexts (Tipton et al., 2016), (2) limited diagnostic information: the item discrimination index does not furnish precise diagnostic details regarding the underlying factors that contribute to the discriminatory performance of a given item. It does not specify whether the item’s difficulty stems from its format, content or other elements. Hence, exclusive reliance on the \( d \)-index could impede the detection of particular item inadequacies and the possibility of improvement (Di Caro et al., 2012), (3) sensitivity to guessing: it is assumed by the item discrimination index that examinees do not attempt to
guess the correct answers (Quaigrain and Arhin, 2017). Nevertheless, in contexts where guessing is probable, such as multiple-choice tests, the index might be susceptible to the impact of guessing behavior and fail to reflect the true discriminatory power of the item accurately; and (4) dependency on item difficulty: Item difficulty impacts the item discrimination index such that items of moderate difficulty tend to have a higher index (Metsämuuronen, 2023). For items that are either extremely simple or extremely difficult, the dependency on item difficulty may hinder the index’s ability to distinguish between high-performing and low-performing individuals.

**RQ3.** What alternatives can be applied to detect undocimologically justified (misfitting) items in the PIE method?

Detecting misfitting items in a test is of the utmost importance in educational and psychological measurement to guarantee the instrument’s validity and reliability. Misfitting items are test items that fail to perform as intended, potentially leading to the distortion of measurement outcomes. A multitude of approaches can be implemented to detect and rectify misfitting items; these approaches are widely utilized within the domain of item response theory (IRT) and qualitative approaches. Utilizing Rasch item fit statistics obtained from IRT models constitutes one method for identifying misfitting items. The relationship between item responses and latent traits is modeled in IRT models, which provide a framework for evaluating item performance. By calculating fit statistics, such as the item fit residual or item fit index, one can assess the degree to which the response pattern of a given item differs from the expected response pattern calculated by the model. Large residuals or significant deviations serve as indicators of possible misfitting items (Metsämuuronen, 2023). Typically used fit indices consist of the infit and outfit mean square statistics (MNSQ) (discrepancy between the observed and expected responses for each item) sensitive to different response patterns and z-standardized (ZSTD) (residuals of the item responses, allowing for comparison across different items) fit indices (Linacre, 2002). The acceptable threshold for MNSQ and ZSTD can be stated as follows: MNSQ values below 0.7 strongly imply overfitting, whereas values above 1.3 indicate underfitting or the potential for misfit. Besides, a threshold of $\pm 2.0$ is widely acknowledged for ZSTD. Indicating a satisfactory model fit, ZSTD values falling within the interval of $-2.0$ to $2.0$ are deemed acceptable (Linacre, 2002).

An alternative approach involves utilizing item characteristic functions (ICFs) or item characteristic curves (ICCs) in IRT for analysis. The relationship between the probability of endorsing an item and the latent trait level is illustrated by these curves (Bean and Bowen, 2021). Through the examination of the ICCs or ICFs’ shape, researchers can detect items that deviate from the expected response. For instance, an item with an unexpected response pattern, such as a reversed response pattern or a floor or ceiling effect, may indicate a misfit. In addition, cognitive interviews and expert judgment are qualitative approaches that can be utilized to identify misfit items. By examining the items’ content and format, expert reviewers can detect potential flaws or ambiguities. Cognitive interviews encompass the process of gathering verbal responses from a sample of examinees regarding their thought processes and comprehension of the test items as they are administered the items (Meadows, 2021). Using these interviews, scholars can acquire valuable knowledge regarding the possible causes of item malfunction, including vague instructions or ambiguous item terminology.

A study by Kockaert and Segers (2017) evaluated a legal text using the PIE method. To evaluate the adequacy of specific items in the text, the study analyzed the $p$ and $d$-values of each item. At the outset, ten items were preselected from the source text (Table 2). They prepared a list of correct and incorrect solutions for ten preselected items among nineteen participants.
The *p*-value was applied to ascertain the acceptability of the items. Items attaining *p*-values between 20 and 90 were deemed adequate by the predetermined criteria. The researchers then calculated the d-index, which was determined to be by the 27% rule after identifying the items that satisfied this criterion. To enhance the selection process, the d-index was employed, which served as an indicator of the items' discrimination. Score calculation was limited to items demonstrating “docimologically justified” characteristics, as indicated by their optimal *p* and d-index values. Items 3, 4, 5, 8 and 9 were determined to satisfy the selection criteria for this study. On the contrary, Rasch Item Fit statistics were subsequently utilized to evaluate the discrimination and difficulty of the items to reach a conclusive decision on their inclusion.

As shown in Table 3, the Rasch item fit statistics revealed that items 1, 2, 4 and 9 were positioned beyond the acceptable threshold based on (ZSTD) (−2 ≤ ZSTD ≤ +2) and

<table>
<thead>
<tr>
<th>Entry number</th>
<th>Total count</th>
<th>JMLE measure S. E</th>
<th>Model S. E</th>
<th>Infit MNSQ ZSTD</th>
<th>Outfit MNSQ ZSTS</th>
<th>Exact OBS%</th>
<th>Match EXP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>−1.01 0.83</td>
<td>1.53 0.96</td>
<td>2.34 1.42</td>
<td>82.4 89.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>0.00 0.63</td>
<td>1.16 0.53</td>
<td>1.66 1.31</td>
<td>82.4 80.1</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>−1.01 0.83</td>
<td>1.32 0.68</td>
<td>0.84 0.10</td>
<td>82.4 89.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>0.68 0.55</td>
<td>1.13 0.60</td>
<td>0.99 0.08</td>
<td>58.8 70.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>0.97 0.54</td>
<td>1.02 0.14</td>
<td>0.98 0.02</td>
<td>64.7 67.0</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>0.36 0.58</td>
<td>0.92 −0.20</td>
<td>0.80 −0.43</td>
<td>76.5 74.8</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>−1.01 0.83</td>
<td>0.79 −0.19</td>
<td>0.87 0.13</td>
<td>94.1 89.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.68 0.55</td>
<td>0.82 −0.75</td>
<td>0.84 −0.44</td>
<td>82.4 70.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>0.00 0.63</td>
<td>0.77 −0.60</td>
<td>0.58 −0.87</td>
<td>82.4 80.1</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>0.36 0.58</td>
<td>0.76 −0.84</td>
<td>0.64 −0.94</td>
<td>76.5 74.8</td>
<td>INVS</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.** Correct and incorrect solutions

**Source(s):** Table courtesy of Kockaert and Segers (2017, p. 156)

The *p*-value was applied to ascertain the acceptability of the items. Items attaining *p*-values between 20 and 90 were deemed adequate by the predetermined criteria. The researchers then calculated the d-index, which was determined to be by the 27% rule after identifying the items that satisfied this criterion. To enhance the selection process, the d-index was employed, which served as an indicator of the items' discrimination. Score calculation was limited to items demonstrating “docimologically justified” characteristics, as indicated by their optimal *p* and d-index values. Items 3, 4, 5, 8 and 9 were determined to satisfy the selection criteria for this study. On the contrary, Rasch Item Fit statistics were subsequently utilized to evaluate the discrimination and difficulty of the items to reach a conclusive decision on their inclusion.

As shown in Table 3, the Rasch item fit statistics revealed that items 1, 2, 4 and 9 were positioned beyond the acceptable threshold based on (ZSTD) (−2 ≤ ZSTD ≤ +2) and

<table>
<thead>
<tr>
<th>Entry number</th>
<th>Total count</th>
<th>JMLE measure S. E</th>
<th>Model S. E</th>
<th>Infit MNSQ ZSTD</th>
<th>Outfit MNSQ ZSTS</th>
<th>Exact OBS%</th>
<th>Match EXP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>−1.01 0.83</td>
<td>1.53 0.96</td>
<td>2.34 1.42</td>
<td>82.4 89.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>0.00 0.63</td>
<td>1.16 0.53</td>
<td>1.66 1.31</td>
<td>82.4 80.1</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>−1.01 0.83</td>
<td>1.32 0.68</td>
<td>0.84 0.10</td>
<td>82.4 89.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>0.68 0.55</td>
<td>1.13 0.60</td>
<td>0.99 0.08</td>
<td>58.8 70.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>0.97 0.54</td>
<td>1.02 0.14</td>
<td>0.98 0.02</td>
<td>64.7 67.0</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>0.36 0.58</td>
<td>0.92 −0.20</td>
<td>0.80 −0.43</td>
<td>76.5 74.8</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>−1.01 0.83</td>
<td>0.79 −0.19</td>
<td>0.87 0.13</td>
<td>94.1 89.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.68 0.55</td>
<td>0.82 −0.75</td>
<td>0.84 −0.44</td>
<td>82.4 70.6</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>0.00 0.63</td>
<td>0.77 −0.60</td>
<td>0.58 −0.87</td>
<td>82.4 80.1</td>
<td>INVS</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>0.36 0.58</td>
<td>0.76 −0.84</td>
<td>0.64 −0.94</td>
<td>76.5 74.8</td>
<td>INVS</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** Items' fit statistics

**Source(s):** Authors’ own work
(0.70 ≤ MNSQ ≤ 1.30) fit indices (Linacre, 2002), respectively. Consequently, including these items in the test was regarded as inappropriate. Poor item construction (e.g., items with multiple translations), item irrelevance (items lacking relevance to the construct being assessed), guessing, item bias (systematic variations in item performance among participants) and local dependencies (items with shared characteristics that exert an influence on one another) were identified as the reasons for their exclusion.

**RQ4.** What alternatives can be applied to check the guessing parameter in the PIE method?

To detect and compute the guessing parameter in item-based approaches such as the PIE method, the probability that test-takers will incorrectly guess the response to a given item must be evaluated. Guessing entails arbitrarily choosing the correct answer, devoid of the requisite expertise or knowledge. It is critical to estimate the guessing parameter precisely to guarantee the reliability and accuracy of item-based assessments. This research paper theoretically proposes several approaches for quantifying the extent of guessing within the PIE method. The PIE method utilizes classical statistical parameters such as $p$ and $d$-values, which date back to the pre-computer era (Eyckmans and Anckaert, 2017, p. 45).

A potential method for estimating the guessing parameter involves the application of IRT models, specifically the two-parameter logistic (2PL) model or the three-parameter logistic (3PL) model (Brwon et al., 2014). These models facilitate the estimation of item parameters, which include the parameters used for guessing. The parameter denoting guessing signifies the likelihood that examinees, notwithstanding their deficiency in the necessary knowledge or abilities, will correctly identify the answer to a given item (Brwon et al., 2014). The estimation of the guessing parameter occurs in conjunction with the item difficulty and discrimination parameters in the 3PL model. The parameter used for guessing is specified as $c$ and is generally limited to a range of values from 0 to 1 (Stenhaug and Domingue, 2022). A lower value indicates a lower probability of guessing, while a higher value suggests a higher probability of guessing.

In contrast, the 2PL model does not incorporate an explicit estimation of the guessing parameter. However, it operates under the assumption that the guessing parameter remains constant at 0.25 and concentrates primarily on estimating item difficulty and discrimination (Brwon et al., 2014). This fixed value is predicated on the assumption that individuals taking the test have a 25% chance of guessing when presented with four response options. The item difficulty parameter, commonly written as $b$, signifies the minimum level of the latent trait a test-taker must possess to have a 50% chance of giving an accurate response. It designates the position along the latent trait continuum at which the item’s difficulty level is neither too easy nor too difficult. A greater value of difficulty indicates that a higher degree of the latent trait is necessary to achieve a successful response on the item. Conversely, a lesser value of difficulty suggests that the item is comparatively less challenging. The item discrimination parameter, commonly represented as $a$, indicates the degree to which the item distinguishes between individuals who possess various degrees of the latent trait (Stenhaug and Domingue, 2022). Higher discrimination values indicate that the item differentiates test-takers with varying abilities with greater efficacy. Conversely, lower discrimination values indicate that the item cannot distinguish between individuals effectively. The 2-PL model offers numerous benefits. This model permits a separate estimation of item difficulty and discrimination, facilitating a more intricate comprehension of item performance (Stenhaug and Domingue, 2022). Additionally, it offers a resilient framework for evaluating item response data, capable of accommodating items in both dichotomous (true/false) and polytomous (multiple-choice) formats (Lipovetsky, 2021).

### 5. Theoretical bottlenecks of the PIE method

As described in the following section, the PIE Method in TE has a number of theoretical bottlenecks. (1) The PIE method does not provide any justification regarding how to segment
a source text for items’ preselection: one bottleneck pertains to the absence of specific instructions or justifications regarding the segmentation of the source text into more manageable segments or items for assessment. It becomes arduous to ascertain the representativeness and coherence of the items under evaluation without adequate justification or criteria for segmentation. To overcome this limitation, scholars may contemplate implementing well-established principles of text segmentation, including an emphasis on cohesive units or significant syntactic structures. Furthermore, the integration of expert judgment or the implementation of studies can assist in ascertaining suitable segmentation standards following specific goals and the features of the TE; (2) The PIE method does not provide any justification regarding the acceptable number of preselected items as per the course content: another bottleneck of the PIE method is the lack of a justification for establishing the ideal number of preselected items, which should be determined by the course content and evaluation objectives. The absence of this justification poses a challenge in guaranteeing that the chosen items sufficiently reflect the abilities or competencies under evaluation. To overcome this constraint, researchers may undertake a comprehensive needs analysis and ensure that the number of pre-selected items corresponds to the intended learning outcomes or evaluation goals. This may entail taking into account variables such as the extent and complexity of the course content, the time allotted for assessment, and the intended degree of accuracy in gauging translation abilities; (3) The PIE method does not provide any justification regarding the appropriate number of evaluators within the research: the number of evaluators employed in the TE process is not adequately justified, which constitutes this bottleneck. An inadequate or excessive number of evaluators may compromise the reliability and generalizability of the evaluation results. To surmount this limitation, scholars may utilize established protocols for ascertaining the ideal number of evaluators, including statistical power analysis. By guaranteeing an adequate number of evaluators, researchers can strengthen the reliability of the evaluation and mitigate potential biases that may arise from the subjectivity of individual evaluators, and (4) The PIE method does not provide any justification regarding the length of the source text for the analysis: one drawback of the PIE method is the lack of a justification for establishing the ideal length of the source text. The length of the source text may impact the complexity, representativeness and diversity of the evaluated translated items. To mitigate this bottleneck, scholars may consider variables that influence the representativeness of the source text, including genre, text type and targeted translation skills. By incorporating established guidelines for text length determination or conducting pilot studies to assess the impact of different text lengths, researchers can establish a justified approach to selecting source texts for PIE analysis.

6. Conclusions
In conclusion, this research paper provided insights into the academic and professional applications of the PIE method, elucidating its capabilities and drawbacks. Despite the potential of the PIE method in evaluating translations, it is advisable to proceed with caution owing to the old-fashioned features of its parameters. Consequently, additional research and enhancement of the PIE method are required, specifically through the application of IRT. The research paper underscored the necessity for further empirical investigation to assess the PIE method’s reliability and validity compared to other well-established methods for evaluating translation, including the holistic, analytic and CDI approaches. This research endeavor would enhance the overall comprehension of the obstacles encountered in the PIE method within an academic context. In addition, the PIE method necessitates reforming statistical measures, including the p-value and d-index, to conform to contemporary statistical and computational equations. Incorporating current methodologies and techniques will augment the PIE method’s scoring approach in accuracy and precision, guaranteeing its relevance and applicability in TE practices.
Note
1. A concomitant variable refers to a variable that deviates from the major focus of research but exhibits potential interaction with the variable(s) under analysis.

References


Corresponding author
Alireza Akbari can be contacted at: dictogloss@gmail.com, alireza.akbari@fgn.ui.ac.ir

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com