Colman, A., Segers, W. & Verplaetse, H. (2021). Refining the PIE method (Preselected Items Evaluation) in translator training. *Current Trends in Translation Teaching and Learning E,* 236 – 276. 8, 236 – 276. https://doi.org/10.51287/cttle20218

REFINING THE PIE METHOD (PRESELECTED ITEMS EVALUATION) IN TRANSLATOR TRAINING

Amy Colman, Winibert Segers and Heidi Verplaetse

KU Leuven

Abstract

This paper explains how to use the PIE method, a criterion- and norm-referenced analytical translation evaluation method, with particular emphasis on translator training. In addition, it sheds light on the test construct and the preparation phase that precedes the PIE evaluation. The source text and item selection, as well as the dichotomous categorisation of translation solutions are discussed in detail. This paper also clarifies and refines two contentious issues in the psychometric component of the PIE method. Firstly, the p-value range, which the authors have revised. Secondly, the calculation of the d-index by means of the extreme groups method. The authors propose two additional methods to calculate the d-index, viz., the adjusted and unadjusted item-total correlation, which, unlike the extreme groups method, take into account all test takers rather than just a set percentage of test takers.

Colman, A., Segers, W. & Verplaetse, H. (2021). Refining the PIE method (Preselected Items Evaluation) in translator training. *Current Trends in Translation Teaching and Learning E,* 236 – 276. 8, 236 – 276. https://doi.org/10.51287/cttle20218

Keywords: PIE, Preselected Items Evaluation, translation evaluation, translator training

1. INTRODUCTION

Preselected Items Evaluation (PIE) is a criterion- and norm-referenced translation evaluation method developed in 2014 by Winibert Segers and Hendrik Kockaert (Kockaert & Segers, 2014). This analytical method (van Egdom et al., 2018a, p. 46) was originally developed for educational purposes, but it can also be used in professional contexts (Segers et al., 2018, p. 42). In fact, PIE has proven its worth as a transparent and objective evaluation method for legal translations in the framework of Qualetra (Quality in Legal Translation) [i] . In addition, PIE offers potential for automation by enabling item-based computerised translation evaluation. It has so far been implemented, albeit in a modified version without the psychometric component and with weighting, in the translation and revision software TranslationQ developed by Televic Education in collaboration with KU Leuven (van Egdom et al., 2018a, p. 35; van Egdom et al., 2018b, p. 34).

The following sections outline the way PIE is used, specifically in translator training, both for continuous assessment and final examinations.

Several aspects of PIE that need further research and refining are discussed, and solutions are proposed for a number of them.

2. PIE: HOW DOES IT WORK?

PIE is based on item analysis, which as McCowan & McCowan (1999, p. 3) explain "uses statistics and expert judgment to evaluate tests based on the quality of individual items, item sets, and entire sets of items, as well as the relationship of each item to other items".

It consists of five phases (Segers et al., 2018, pp. 41-42; Tijtgat & Segers, 2019, pp. 322-323):

(1) Item selection: selection of the items to be evaluated
(2) Dichotomous categorisation of translation solutions for each item as correct or incorrect
(3) Calculation of the test takers' raw scores
(4) Item analysis: calculation of the difficulty (p-value) and discrimination index (d-index) of each item
(5) Recalculation of the test takers' scores

The first two phases are criterion-referenced. They are followed by the calculation of the raw scores. The final two phases are norm-referenced. A

criterion-referenced method evaluates scores based on specific criteria, e.g., "what the examinees know or what they can do" (Reynolds et al., 2006, p. 60). A norm-referenced method compares the scores of the test takers with those of a reference group (*ibid*., p. 59). In the case of PIE, the reference group is made up of the group of test takers themselves rather than a different group that acts as the benchmark (Tijtgat & Segers, 2019, p. 323).

PIE can be used as a criterion-referenced method only. In that case, the evaluation stops at phase three. If the evaluators wish to include the norm-referenced component, they should also carry out the psychometric analysis in phase four and five. In the literature, the criterion-referenced version of PIE is labelled "PIE Light" (Kockaert & Segers, 2017, p. 153; Tijtgat & Segers, 2019, p. 320). As 'Light' carries a somewhat pejorative connotation, the authors propose the term *PIE criterion-referenced only*. The authors also propose the label *PIE criterion- and norm-referenced* for the full version of PIE.

It must be stressed that the raw scores in *PIE criterion-referenced only* may differ from those in *PIE criterion- and norm-referenced*. The reason for this possible discrepancy lies in the fact that the latter uses psychometric analysis, viz. the

calculation of the difficulty (p-value) and discrimination index (d-index) of each preselected item. As such, *PIE criterion- and norm-referenced* takes into consideration the performance of the entire group of test takers, while *PIE criterion-referenced only* focuses exclusively on the scores of the individual test takers.

Before providing a detailed explanation of these five phases, it is important to focus on the design of the PIE test, particularly on the source text selection. This is essential, as it influences the item selection and analysis, and hence the outcome of the evaluation. In addition, in the preparatory phase, the role of the evaluator – or preferably evaluators – should be clearly defined to guarantee the highest possible degree of inter-rater and intra-rater reliability[ii].

3. PREPARATION: SOURCE TEXT SELECTION

The literature on PIE does not discuss the selection and characteristics of the source text (Kockaert & Segers, 2014; Segers & Kockaert, 2016; Kockaert & Segers, 2017; Segers et al., 2018; Tijtgat & Segers, 2019). However, these are crucial factors that influence the effectiveness of the PIE analysis and

the outcome of the evaluation. They therefore merit further attention.

PIE is a method that relies on the analysis of preselected items chosen in a criterion-referenced manner. Therefore, the source text should cover a relevant topic and contain items that are worth testing. The source texts and preselected items in translator training should be closely linked to the course contents. In addition, they should enable testing of the intended learning outcomes and the competences to be achieved (van Egdom et al., 2018a, p. 36). In the framework of a course that focuses on the translation of business texts, for example, the lecturer may select a source text about shipping or trade dense in specialised terminology taught during the course. In order to test whether the students have assimilated this terminology, the lecturer may opt for preselected items from this item category only. However, it may also be useful to choose a source text that allows for a varied item selection. If a text contains mostly terminological items, for example, it might not give an insight into the students' ability to translate grammatical structures. It will merely prove whether or not they are able to study terminology and/or correctly use glossaries or dictionaries. If the focus of the evaluation is a specific item category, such as

business terminology, it is advisable to select a source text that enables this targeted evaluation.

It is important to note that PIE can be used for continuous assessment and formative purposes as well as for final examinations and summative purposes. Students and lecturers can use PIE *criterion-referenced only* and/or PIE *criterion- and norm-referenced* to create a portfolio of translations, which enables them to keep track of the scores and the errors made. However, as PIE does not include an error categorisation method, ideally it is combined with other evaluation methods which allow for the errors made to be categorised, which is especially useful for formative purposes and course design. An example are the error categories of the ATA Framework for Standardized Error Marking[iii], which are accompanied by a detailed description. This provides a clearer insight into the linguistic and translational error categories and hence remedial measures.

Another key point to take into consideration is that the selected topic or text may result in a biased test design. Boyle and Fisher (2007, p. 19) explain that any factors that may cause bias should be removed or limited. An example of test bias would be the selection of a Dutch source text containing typically Flemish words to be translated into English by a

group of Flemish and Dutch students. The test bias arises from the fact that the Flemish students would be at an advantage compared to their Dutch peers. It is, of course, impossible to fully exclude test bias, but known factors must be taken into consideration. A biased test design might result in a flawed PIE evaluation because certain items might be far easier to translate for some test takers than for others. This would affect the p-value and d-index of the items in question, as discussed in more detail in section 4.4.

Ideally, the source text should be chosen by more than one evaluator. In practice, this is, of course, not always possible due to limited timeframes or resources. However, if possible, an intersubjective consensus should be reached on the source text to be used. Likewise, as explained in section 4.1., the items to be evaluated should also be chosen based on an intersubjective consensus. The same goes for the dichotomous categorisation of correct and incorrect translation solutions for each item. This is thought to increase the inter-rater reliability of PIE, but empirical research is needed to substantiate this theory.

As for the source text length, the literature on PIE suggests selecting a source text of approximately 200 words to be translated in 90 minutes (Kockaert & Segers, 2014, pp. 243-244; Segers & Kockaert,

2016, pp. 70-71; Kockaert & Segers, 2017, p. 156; Segers et al., 2018, p. 43). This applies to both translation tests and texts translated during translation workshops. In translator training, the length of the source text is often limited (Segers, 2007, p. 21). O'Brien (2009, pp. 261-262) explains that the source texts used in translation experiments usually have a length of 200 to 300 words, which makes them shorter than the texts generally translated by professional translators. However, many translation tests for professional translators are around 250 words in length. This is the "hypothetical norm" (Koby & Baer, 2005, p. 42) used by the American Translators Association (ATA). The ATA certification exam consists of two texts of "225 to 275 words each" to be completed in three hours[iv]. Thus the source text length and the timeframe allocated for PIE tests are in accordance with standard practice in translator training and professional testing.

When designing the PIE translation test, lecturers must ensure that the source text is not excessively long. PIE has so far not been tested with longer source texts, neither in a classroom setting nor during examinations. While not validated empirically yet, it is thought that an excessively long source text could result in some students attempting to translate it in full, but carelessly, while others may

translate only part of it, but very meticulously. This would affect the construct validity of the evaluation. Williams (2009, p. 5) defines construct validity as "the extent to which an evaluation measures what it is designed to measure, such as translation skills". An excessively long translation test could end up measuring stamina rather than translation skills. A shorter source text that covers the curriculum perfectly and tests the students' ability to achieve the intended learning outcomes demonstrates both construct validity and content validity, viz., "the extent to which an evaluation covers the skills necessary for performance" (*ibid*.). It must be highlighted that the validity of PIE has not yet been researched in empirical studies. As a result, these are merely theoretical assumptions. Validity studies are very complex in nature, not in the least because there are many different types of validity that can be tested[v]. In this context, the translation skills referred to above are difficult to define. Translation tests are generally accepted as a tool to evaluate whether the students have the required competence to translate from the source language into the target language. Therefore, construct validity is rarely questioned in this field (Eyckmans & Anckaert, 2017, p. 42). However, a clear definition of competence is required, which is no easy feat, as there is no consensus among scholars (Arango-Keeth & Koby, 2003, p. 119).

In addition, there is a lack of psychometric testing in this field (Eyckmans et al., 2013, p. 1). PIE tests, with their psychometric component, might help bridge this gap. It must be highlighted, however, that the number of test takers within a single translator training institution is often too limited to obtain a representative sample for psychometric testing. Therefore, collaborations between institutions are encouraged to set up large-scale empirical studies into the validity of PIE. An example is the experiment set up by van Egdom et al. (2018a), which is discussed in more detail in section 4.1.

Once the source text has been selected, we can move on to the item selection, which is described below.

4. THE FIVE PHASES OF PIE

4.1. Item selection

The literature on PIE suggests that 10 items are selected in a source text of approximately 200 words (Kockaert & Segers, 2014, p. 244; Segers & Kockaert, 2016, p. 71; Kockaert & Segers, 2017, p. 156; Segers et al., 2018, p. 43). When the PIE method was initially developed in 2014, this set number of 10 items was randomly chosen to allow for a time-efficient evaluation. The aim was to

enable lecturers to consistently evaluate large numbers of translations within a limited timeframe, and assign a score out of ten. It must be noted that so far, no empirical research has been carried out into the ideal number of items for PIE evaluations. It is also important to stress that the test takers should not be informed of the items that have been preselected (van Egdom et al., 2018b, p. 35). However, it is recommended that they are informed of the fact that the evaluation will be based on a number of preselected items only.

The preselected items are short parts of the source text on which the translation evaluation will be based, for example words or phrases (Kockaert & Segers, 2014, p. 238). This segmentation is very important. Entire paragraphs are too long to be used as items. The same goes for entire sentences or long phrases. After all, it would be too difficult, if not impossible, to categorise their translations dichotomously, as correct or incorrect. Proper segmentation enables the evaluators to isolate specific items and simplifies their work: a partially correct translation of a long item would require evaluators to decide whether or not to consider it an error as a whole. They would also be unable to get a conclusive insight into the issues the students faced. Did they struggle to translate the verb tense? Or maybe they failed to translate the terminology

correctly?

As for the item categories, as explained in section 3, in translator training, the items are chosen based on the curriculum and the intended learning outcomes (Segers et al., 2018, p. 43; Segers & van Egdom, 2018, p. 80). Examples include legal terminology in courses that focus on the translation of legal texts, or cultural references and/or verb tenses in literary translation courses. In the recruitment of professional translators, they are linked to the desired profile (*ibid*.), for example a translator specialised in technical texts. It must be noted that certain item categories do not lend themselves well to a PIE analysis. The translations of idioms and other creative language usage may be difficult to categorise as correct or incorrect because there are too many possible correct translations. If the evaluators unexpectedly find that the test takers proposed a correct translation that was not included in the list of correct solutions compiled before the analysis, it can be added to that initial list (Kockaert & Segers, 2017, p. 153). However, the evaluators then need to go through all the translations again to verify whether other test takers came up with that same unexpectedly correct solution (van Egdom et al., 2018a, p. 37). If several evaluators were involved in the item selection, ideally they should approve this unexpected

translation solution too, which may be time-consuming. This is a limitation of PIE.

Like the source text, ideally the items should also be selected by multiple evaluators, based on an intersubjective consensus. This is not explicitly mentioned in the literature, although when discussing the item selection, Kockaert and Segers (2014, p. 238) do use the plural of 'evaluator'. An item selection by more than one evaluator – two or more lecturers of the same course, for example – safeguards the inter-rater reliability. If the different evaluators disagree on an item, it should be discarded.

The inter-rater reliability of the item selection in PIE was tested by van Egdom et al. (2018a, p. 39). The researchers asked multiple evaluators from three different translation programmes to select a number of items in a source text jointly chosen by two of them. Despite differences in the focus of the translation programmes – two being academic and one vocational – and the different levels of experience of the participating students [vi], the evaluators selected similar items. The authors of this experiment agree that "the similarities seem to suggest that a clear translation brief and a certain level of teaching and evaluation experience effectively do provide us with the hoped-for

operational measures for TT evaluation and, as a consequence, do add to the construct validity of PIE testing" (*ibid*., p. 38-40). Nevertheless, they acknowledge that further research is required to determine the 'ideal' number of evaluators needed for a "valid" item selection (*ibid*., p. 48). This also calls for agreement on the exact definition of "valid" in this context.

As van Egdom et al. (2018a, p. 48) rightly point out, it is also unclear how many evaluators are required for a valid item selection and what their expertise level should be. In addition, as with the source text selection, it must be acknowledged that from a practical point of view, item selection by multiple evaluators may not be feasible due to limited resources and/or timeframes.

Once the items have been preselected, we can move on to the dichotomous categorisation of translation solutions.

4.2. Dichotomous categorisation of translation solutions

In this phase, the evaluators list all the possible translation solutions for each item, labelling them as correct or incorrect. This is referred to as a 'dichotomous' categorisation of errors. PIE is a

dichotomous method because it distinguishes between correct and incorrect solutions only (Kockaert & Segers, 2017, pp. 150-151). Errors are not weighted, viz., they are not classified as major, average and minor (Segers et al., 2018, p. 42), neither are they assigned bonus points (Eyckmans & Anckaert, 2017, p. 44). Just as for the item selection, if possible, this phase of PIE relies on an inter-subjective consensus between at least two evaluators. Despite the lack of extensive research into this aspect of PIE, it can safely be assumed that the evaluation based on a set number of items as well as the categorisation of possible translation solutions as correct or incorrect prior to the PIE analysis guarantee a high level of inter- and intra-rater reliability. After all, if two evaluators carry out the same analysis based on the same preselected items and dichotomous categorisation of correct and incorrect translation solutions, any discrepancies between their scores can only be attributed to human error.

There is no "grey area" because the evaluators establish in advance which translation solutions are correct and which ones are incorrect (Segers et al., 2018, p. 42). This does not mean the list of correct and incorrect solutions is exhaustive. If the evaluators unexpectedly come across additional correct alternatives, they are included in the list a

posteriori and all tests are evaluated once again (Kockaert & Segers, 2017, p. 153; van Egdom et al., 2018a, pp. 36-37). Ideally, if multiple evaluators were involved in the item selection and dichotomous categorisation of possible translation solutions, they should also be consulted prior to accepting an unexpected solution as correct.

It could be argued that this dichotomous categorisation is too rigid because language is dynamic and not every solution is necessarily strictly right or wrong. In addition, one evaluator may consider an error minor, while another may consider it a major error. However, if there were a grey area or a weighting component, it would be impossible to carry out a psychometric analysis using statistical formulas, viz., the calculation of the item difficulty (p-value) and discrimination index (d-index) of each preselected item. This does raise the question of what evaluators should do when faced with translation solutions that are neither entirely correct nor entirely incorrect. When using PIE, the evaluators should only choose items with translation solutions that can unequivocally be categorised as correct or incorrect. When in doubt, they should rely on authoritative sources to check whether a solution can be considered correct or incorrect, such as institutions that develop national language policies. The expertise of the evaluators is

also essential in this view, as is the translation brief. If the target text is aimed at an American audience, for example, only American English translation solutions should be accepted as correct. Their British English equivalent should be categorised as incorrect. The evaluators must also ensure that there is no scope for doubt. In translator training, contentious items could lead to disputes, even of a legal nature if they involve grading. The evaluators should therefore always be able to present a list of references to justify their decisions. Merely relying on the number of hits in a search engine should not be regarded as proof of correctness. It could therefore be useful to draw up a list of authoritative sources to be used for the evaluation in advance.

These issues show that the item selection should not be taken lightly and evaluators should leave no scope for doubt in their dichotomous categorisation of possible translation solutions. Once the item selection has been concluded and all evaluators involved are satisfied with the outcome, we can start calculating the test takers' raw scores, as described below.

4.3. Calculation of the raw scores

The first two phases of PIE, as described above, are criterion-referenced, viz., (1) the selection of the

items to be evaluated and (2) the dichotomous categorisation of the correct and incorrect translation solutions for each preselected item (Tijtgat & Segers, 2019, p. 322).

According to van Egdom et al. (2018a, p. 34), "the assessment and evaluation of learning outcomes and competences seem to call for criterion-based testing methods". The authors explain that Calibration of Dichotomous Items (CDI)[vii], the predecessor to PIE, is norm-referenced only, which compromises its construct validity. With this in mind, PIE was developed to include both a criterion-referenced and a norm-referenced component. However, evaluators can opt to stop the analysis after the third phase of PIE, viz., the calculation of the raw scores. In that case, they have used *PIE criterion-referenced only*.

The raw scores are the test takers' scores before quantitative or qualitative analysis (Zedeck, 2014, p. 296). They are calculated by deducting the incorrectly translated items from the total number of items. An example: if a test taker incorrectly translated four items out of ten, their raw score is six out of ten, as illustrated in Table 1.

Table 1. Calculation of the raw scores

Number of preselected items	Correctly translated items	Incorrectly translated items	Raw score
10	6	4	6/10

Evaluators should be aware of the fact that these raw scores may differ considerably from those obtained in the full version of PIE, viz., *PIE criterion- and norm-referenced*. The latter are the recalculated scores following the psychometric analysis. The reason for this possible discrepancy is that the raw scores do not take into consideration the performance of the entire group of test takers, but only that of the individual test takers.

If the evaluators wish to do so, they may move on to the fourth and fifth phases, which constitute the psychometric analysis, viz., the calculation of the item difficulty (p-value) and discrimination index (d-index). In these norm-referenced phases the scores are interpreted based on the performance of the tested group (Bachman, 2004, p. 30). This norm-referenced component is important in an educational context because it allows the evaluators to distinguish between test takers and check whether the curriculum has been assimilated. Norm-referenced tests can also help refine the curriculum. If, for example, after the first semester it turns out

that certain item categories are problematic for a large number of students, the lecturer can adapt the curriculum for the remainder of the year to address these issues.

To date, there are no guidelines on the effective sample size for PIE analyses. However, it must be stressed that *PIE criterion- and norm-referenced* is not recommended for excessively small groups of students.

4.4. Psychometric analysis

The next two phases of PIE are the calculation of the item difficulty (p-value) and the discrimination index (d-index) of each item.

4.4.1. Item difficulty (p-value)

The item difficulty or p-value [viii] indicates how many test takers have translated an item correctly. The higher the p-value, the easier the item. Since this definition is somewhat confusing, the p-value is sometimes referred to as "facility value" (Pidgeon & Yates, 1968, p. 57), "index of facility" (Ebel, 1979, p. 263) or "item facility" (Weir, 2005, p. 202). In this paper, the p-value is referred to as *item difficulty*, in accordance with the literature on PIE. The p-value is calculated by dividing the

number of correct translation solutions by the total number of test takers, and ranges from 0 to 1 (Reynolds et al., 2006, p. 143; McCowan & McCowan, 1999, p. 18).

No consensus has been reached on the 'ideal' p-value when using PIE. In the literature, a good p-value lies between 0.27 (too easy) and 0.79 (too difficult) (Segers and Kockaert, 2016: 74; Segers et al., 2018, p. 47; van Egdom et al., 2018a, p. 37). This is the most commonly quoted p-value in the literature on PIE, but some sources use a different range, viz., 0.20 to 0.90 (Kockaert & Segers, 2014, p. 245; Eyckmans & Anckaert, 2017, p. 44).

The most commonly used p-value for PIE analyses, viz., 0.27-0.79, appears to have been deducted from the extreme groups method, which is used to calculate the discrimination index (d-index) of each item, a measure to distinguish between strong and weak test takers, which is explained in further detail in section 4.4.2.1.

The extreme groups method only takes into consideration 54% of all test takers, viz., the 27% of test takers who obtained the highest scores and the 27% of test takers who obtained the lowest scores (Segers et al., 2018, p. 48). The authors of the PIE

method calculate the item difficulty (p-value) range based on the extreme groups method. 0.27 refers to the weakest group or bottom group, i.e. the group of test takers who obtained the lowest scores. The strongest group or top group, i.e. the group that obtained the highest scores, is then calculated as follows:

Total test takers = 100% - 27% = 73% = 0.73

Hence, the value 0.79 that is cited in the literature on PIE is incorrect and should be corrected as 0.73.

It is open to question whether the item difficulty (p-value) can be calculated using the same methods as those used to calculate the discrimination index (d-index), in this case the extreme groups method. McCowan and McCowan (1999, p. 18) explain that the "item difficulty is a characteristic of the item and the sample that takes the test". The test takers do indeed to some degree determine the item difficulty. Let's suppose a translation test for students in the final year of translator training were presented to first-year students. The preselected items could include specific terminology covered during the first semester of the final year. It would be reasonable to assume that the difficulty of these items would be judged differently by both groups, with the first-year group struggling more than the final-year group.

However, within one and the same group, particularly when working with small sample sizes, we cannot say with certainty that there is a link between the item difficulty and the discrimination index. Other factors may be responsible for the score differences between strong and weak test takers, such as access to or use of reference sources.

The authors propose to use a p-value range that has not yet been used with PIE but is often quoted in other educational measurement sources (e.g. Allen & Yen, 1978, p. 121) and is 'neater', viz., 0.30 to 0.70. These sources state that ideally, all items combined should have a mean of 0.50 if the scores present a normal distribution.

Table 2. Example calculation of the item difficulty

Total number of test takers	Total number of correct translation solutions for item A	p-value of item A
15	8	8/15 = 0.53

It is important to acknowledge that item difficulty is closely linked to text difficulty. The latter cannot be determined in advance, particularly when comparing different source and target languages. Readability tools can be used to get an insight into the difficulty of the source text. Pavlovic (2007, p. 61) explains that "texts can be compared in terms of

various parameters, such as readability, word frequency, length or genre". She warns that, "problems arise, however, when these tools (which are freely available on the Internet) are used to compare texts written in different languages, as is the case in studies involving directionality of translation" (*ibid*.).

4.4.2. Discrimination index (d-index)

The discrimination index (d-index) shows to which degree each item discriminates between strong and weak test takers. Students who achieved a high (raw) score on the entire test would be expected to have translated an item correctly and students who achieved a low (raw) score overall, would be expected to have translated it incorrectly, not vice versa. The d-index helps us check whether or not that is the case (Weir, 2005, p. 202).

There are different ways to calculate the d-index. The PIE method usually relies on the extreme groups method (Segers et al., 2018, p. 48). Once all the raw scores have been calculated, they should be ranked from high to low. Subsequently, the test takers should be divided into three groups: the top group, the middle group and the bottom group. The middle group is not taken into consideration. The top group includes the students with the highest (raw)

scores on the entire test and the bottom group includes those with the lowest (raw) scores (Ebel & Frisbie, 1991, p. 225). Each group represents a set percentage of the total test population. Ebel (1979, p. 260) suggests 27%, because that "provides the best compromise between two desirable but inconsistent aims: (1) to make the extreme groups as large as possible and (2) to make the extreme groups as different as possible". The PIE method uses this 27% as a reference value for the top and bottom group (Kockaert & Segers, 2017, p. 152).

To calculate the d-index of the preselected items in PIE, the extreme groups method is applied by deducting, for each item, the p-value in the bottom group from the p-value in the top group (*ibid*.). As such, a link is established between the d-index and the p-value of each item. Items should have a discrimination index of at least 0.3 to be included in the calculation of the final test score (Ebel, 1979, p. 267).

Another method to determine the d-index is to calculate the item-total correlation, viz., the correlation between a specific item and the total test score, either adjusted or unadjusted. Adjusted means that the item in question is omitted from the total test score. Unadjusted means that the item is included in

the total test score. A high item-total correlation indicates that the item discriminates well between high-performing and low-performing test takers (Reynolds et al., 2006, pp. 147-148).

The calculation of the item-total correlation based on the unadjusted total test score results in the r_{it} value, viz., the correlation between a specific item (i) and the total test score (t). The calculation of the item-total correlation based on the adjusted total test score results in the r_{ir} value, which indicates the correlation between an item (i) and the rest score (r), viz., the total test score minus the item in question (de Gruijter & van der Kamp, 2008, p. 93).

If the item-total correlation is calculated based on the adjusted total test score, it will be lower than when calculated based on the unadjusted total test score. The reason for this lower value is that the item in question is not included in the total test score and can therefore not cause contamination or inflation of the correlation. Since the r_{it} value takes the item in question into consideration, it is less reliable than the r_{ir} value (Reynolds et al., 2006, pp. 147-148).

4.4.2.1. Discrimination index: example calculations

Below you will find a simple example illustrating the three methods for the calculation of the d-index. As explained earlier, the psychometric analysis is not recommended for small groups of test takers. The example below is therefore intended for illustrative purposes only.

Table 3. Test example: raw scores of 3 students on a test with 5 items (item A, B, C, D and E)

	Item A	Item B	Item C	Item D	Item E	Total raw score /5
Student 1	0	1	1	1	1	4
Student 2	0	1	0	1	0	2
Student 3	1	0	0	0	0	1

1 indicates a correctly translated item
0 indicates an incorrectly translated item

Extreme groups method

Top group = student with the highest score
= student 1 (score: 4/5)
Bottom group = student with the lowest score
= student 3 (score: 1/5)

p-value of item A in the top group = 0/1* = 0
p-value of item A in the bottom group = 1/1 = 1

p-value top group – p-value bottom group = 0-1 = -1

Result: the d-index of item A = -1

* score for item A (in this case 0) divided by the number of students in the group (in this case 1)

Unadjusted item-total correlation (r_{it}) of item A

To calculate the r_{it} of item A using the =CORREL formula in Microsoft Excel, select the 'Item A' column (the scores for item A) as array1 and select the 'Total raw score' column as array2.

Result: the r_{it} of item A is -0.76.

Adjusted item-total correlation (r_{ir}) of item A

To calculate the r_{ir} of item A using the =CORREL formula in Microsoft Excel, select the 'Item A' column (the scores for item A) as array1 and select the 'Total raw score minus score for item A' column as array2 (see Table 4).

Table 4. Test example: calculation of the adjusted item-total correlation (r_{ir}) of item A

	Item A	Item B	Item C	Item D	Item E	Total raw score /5	Total raw score minus score for item A
Student 1	0	1	1	1	1	4	4
Student 2	0	1	0	1	0	2	2
Student 3	1	0	0	0	0	1	0

Result: the r_{ir} of item A is -0.87.

In the example above, the extreme groups method results in a d-index of -1. The r_{it} is -0.76 and the r_{ir} is -0.87. This means that there is a negative correlation, viz., the student who achieved the highest total test score (student 1) did not score well on item A. The student who achieved the lowest total test score (student 3), on the other hand, scored well on item A.

PIE uses the most common d-index reference values in educational testing, namely the Ebel ranges (Ebel, 1979, p. 267), as illustrated in Table 5.

Table 5. Discrimination index value ranges

Discrimination index (d-index)	**Item evaluation**
0.40 and up	Very good items
0.30 to 0.39	Reasonably good items but possibly subject to improvement
0.20 to 0.29	Marginal items, usually needing and being subject to improvement
Below 0.19	Poor items, to be rejected or improved by revision

Ebel suggests to improve items when they do not fall within the desired ranges. With PIE, manipulation of items – and therefore possibly also of the source text – is not recommended. The only improvement or manipulation that is acceptable is resegmentation, viz., selecting a shorter or longer section of the source text as an item.

In the example above, we can conclude that item A should be discarded because it does not discriminate well between test takers. McCowan and McCowan (1999, p. 21) explain that an item with a high d-index discriminates in favour of the test takers with the highest scores. This top group is expected to answer more items correctly. If a large proportion of test takers in the bottom group correctly translate an item, the d-index of that item will be negative, viz. below zero, which means the item is probably

defective (*ibid*.). This may be the result of the bottom group making a correct guess and/or the top group overthinking and interpreting the item incorrectly (*ibid*.).

We now have all the data we need to calculate the test takers' final scores in the last phase of the PIE analysis.

4.5. Recalculation of the raw scores

In the final phase of PIE, the items with a p-value and d-index that do not fall within the established ranges are discarded. The test takers' scores are then recalculated, which may result in major discrepancies between the raw scores and the final scores (Kockaert & Segers, 2017, p. 160). This makes PIE a calibration method, which entails that the evaluator(s) must analyse and adjust the accuracy of the measurement tool (*ibid*., pp. 150-151). Concretely this means that when carrying out the psychometric analysis, the test takers' final scores depend on the calculation of the item difficulty (p-value) and the discrimination index (d-index) of each preselected item.

If too many items are discarded, Van Egdom et al. (2018a, p. 37) argue that the quality of test construct may be at stake. In her PhD research, the first author

of this paper intends to compare a PIE analysis with 10 items on the one hand, and with 20 items on the other to check the possible impact on the final scores.

5. CONCLUSION

This paper highlights that, despite being a promising method for more objective translation evaluation, the PIE method needs to be further studied and refined through empirical research. The authors emphasised the importance of the source text and item selection, and the dichotomous categorisation of translation solutions. An intersubjective consensus should be reached in these phases of PIE so as to achieve a high inter-rater and intra-rater reliability. In addition, it must be analysed whether selecting more than the 10 items prescribed by the literature may generate a different outcome. These aspects had never before received much attention, if any, in the literature. The authors also corrected the p-value range used so far and proposed two additional methods for the calculation of the d-index. Many questions remain unanswered, including the impact of the source text selection on the evaluation, the number of preselected items required and the number of evaluators needed to reach an intersubjective consensus. It is hoped that this paper

will mark the first step towards more extensive empirical research into PIE.

Colman, A., Segers, W. & Verplaetse, H. (2021). Refining the PIE method (Preselected Items Evaluation) in translator training. *Current Trends in Translation Teaching and Learning E,* 236 – 276. 8, 236 – 276. https://doi.org/10.51287/cttle20218

REFERENCES

Allen, M. J. & Yen W. M. (1979). *Introduction to measurement theory.* Brooks/Cole.

Arango-Keeth, F. & Koby, G. S. (2003). Assessing assessment: Translator training evaluation and the needs of industry quality assessment. In B. J. Baer & G. S. Koby (Eds.). *Beyond the Ivory Tower* (pp. 112-134). John Benjamins.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Boyle, J. & Fisher, S. (2007). *Educational testing: A competence-based approach.* Blackwell Publishing.

de Gruijter, D. N.M. & van der Kamp, L. J. T. (2008). *Statistical Test Theory for the Behavioral Sciences*. Chapman and Hall/CRC.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th ed.). Prentice-Hall.

Ebel, R. L. (1979). *Essentials of Educational Measurement.* Prentice Hall.

Eyckmans, J., Anckaert P. & Segers, W. (2009). The perks of norm-referenced translation evaluation. In C. V. Angelelli & H. E. Jacobson (Eds.). *Testing and assessment in translation and interpreting studies* (vol. 14) (pp. 73-93). John Benjamins.

Eyckmans, J., Anckaert, P. & Segers, W. (2013). Assessing translation competence. *Actualizaciones En Communicación Social, 2*, 513-515.

Eyckmans, J. & Anckaert, P. (2017). Item-based Assessment of Translation Competence: Chimera of Objectivity versus Prospect of Reliable Measurement. *Linguistica Antverpiensia New Series-Themes In Translation Studies 16*, 40-56.

Koby, G. S. & Baer, B. J. (2005). From professional certification to the translator training classroom: Adapting the ATA error marking scale. *Translation Watch Quarterly, 1*(1), 33-45.

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

Kockaert, H. & Segers, W. (2017). Evaluation of legal translations: PIE method (Preselected Items

Evaluation). *Journal Of Specialised Translation, 27*, 148-163.

McCowan, R. J. & McCowan, S. C. (1999). *Item Analysis for Criterion-Referenced Tests.* Center for Development of Human Services.

O'Brien, S. (2009). Eye tracking in translation process research: methodological challenges and solutions. In I. M. Mees, F. Alves & S. Göpferich (Eds.). *Methodology, technology and innovation in translation process research: a tribute to Arnt Lykke Jakobsen. Copenhagen studies in language, 38* (pp. 251-266). Samfundslitteratur.

Pavlovic, N. (2007). *Directionality in Collaborative Translation Processes* (Publication No. T.2197-2007) [Doctoral dissertation, Universitat Rovira i Virgili]. http://hdl.handle.net/10803/8770

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

Reckase, M. D. (2009), *Multidimensional Item Response Theory*. Springer.

Reynolds, C. R., Livingston, R. B. & Willson, V. (2006). *Measurement and assessment in education.* Allyn & Bacon/Pearson.

Segers, W. (2007). IJkpuntenmethode. In C. Van de Poel & W. Segers. *Vertalingen objectief evalueren: Matrices en ijkpunten* (pp. 21-25). Acco.

Segers, W. & Kockaert, H. (2016). Can subjectivity be avoided in translation evaluation?. In M. Thelen, G.-W. van Egdom, D. Verbeeck & B. Lewandowska-Tomaszczyk (Eds.), *Łódź Studies in Language, vol: 41, Translation and Meaning: New Series* (pp. 69-78). Peter Lang.

Segers, W. & van Egdom, G.-W. (2018). *De kwaliteit van vertalingen: Een terminologie van de vertaalevaluatie.* Pelckmans Pro.

Segers, W., Kockaert, H. & Wylin, B. (2018). Vertaalevaluatie en subjectiviteit. *Tijdschrift N/f (13)*, 41-51.

Steurs, F., Segers, W. & Kockaert, H. (2015, September 9-11). *Translation Expert (TranslationQ & RevisionQ): Automated Translation Process with Real-time Feedback & Evaluation/ Revision with PIE* [Conference session]. Talking to the World, University of Newcastle, United Kingdom. https://tinyurl.com/snjnseby

Tijtgat, E. & Segers, W. (2019). Wat is een goede vertaling? Vertaalevaluatie: methodes en technieken. In G. De Sutter & I. Delaere, *In balans. Een inleiding tot vertaal- en tolkwetenschap* (pp. 307-328). Acco.

van Egdom, G.-W., Verplaetse, H., Schrijver, I., Kockaert, H., Segers, W., Pauwels, J., Bloemen, H. & Wylin, B. (2018a). How to Put the Translation Test to the Test? On Preselected Items Evaluation and Perturbation. In E. Huertas-Barros, S. Vandepitte & E. Iglesias-Fernández (Eds.), *Quality Assurance and Assessment Practices in Translation and Interpreting* (pp. 26-56). IGI Global. https://doi.org/10.4018/978-1-5225-5225-3.ch002

van Egdom, G.-W., Segers, W., Bloemen, H., Kockaert, H. & Wylin, B. (2018b). Revising and Evaluating with TranslationQ. *Bayt Al-Hikma: Journal for Translation Studies, 2018*(2), 25-56.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Palgrave Macmillan.

Williams, M. (2009). Translation Quality Assessment. *Mutatis Mutandis, 2*(1), 3-23.

Colman, A., Segers, W. & Verplaetse, H. (2021). Refining the PIE method (Preselected Items Evaluation) in translator training. *Current Trends in Translation Teaching and Learning E,* 236 – 276. 8, 236 – 276. https://doi.org/10.51287/cttle20218

Zedeck, S. (Ed.). (2014). *APA dictionary of statistics and research methods.* American Psychological Association.

Endnotes

[i] JUST/2011/JPEN/AG/2975 (Steurs et al., 2015, p. 1).

[ii] Inter-rater reliability: "degree to which two or more evaluators produce the same results in measuring the same characteristics" (van Egdom et al, 2018a, p. 54).

Intra-rater reliability: "degree to which one and the same evaluator produces the same results in measuring the same characteristic" (*ibid.*).

[iii] The error categories only, as found on https://www.atanet.org/certification/how-the-exam-is-graded/error-categories/ (last viewed on 13 October 2021)

[iv] https://www.atanet.org/certification/aboutexams_overview.php (last viewed on 13 October 2021).

[v] Suggested titles for further reading on validity: Weir (2005) and Williams (2009).

Colman, A., Segers, W. & Verplaetse, H. (2021). Refining the PIE method (Preselected Items Evaluation) in translator training. *Current Trends in Translation Teaching and Learning E,* 236 – 276. 8, 236 – 276. https://doi.org/10.51287/cttle20218

[vi] Two groups at third-year BA level – one with an academic focus and the other with a vocational focus – and one group at academic MA level.

[vii] Suggested further reading on CDI: Eyckmans et al. (2009); Eyckmans and Anckaert (2017).

[viii] The letter *p* in p-value stands for the *proportion* of test takers who have provided a correct answer (Reckase, 2009, p. 26).