Eyckmans, J., & Anckaert, Ph. (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.

Item-based assessment of translation competence: Chimera of objectivity versus prospect of reliable measurement

June Eyckmans

Ghent University, Belgium
June.eyckmans@ugent.be

Philippe Anckaert

Université libre de Bruxelles, Belgium
Philippe.anckaert@ulb.ac.be

In the course of the past decade, scholars in Translation Studies have repeatedly expressed the need for more empirical research on translation assessment. Notwithstanding the many pleas for "objectivity" that have been voiced in the literature, the issue of reliability remains unaddressed. Although there is no consensus on the best method for measuring the quality of human or machine translations, it is clear that in both cases measurement error will need to be accounted for. This is especially the case in high-stake situations such as assessments that lead to translation competence being certified. In this article we focus on the summative assessment of translation competence in an educational context. We explore the psychometric quality of two assessment methods: the CDI method (Eyckmans, Anckaert, & Segers, 2009) and the PIE method (Kockaert & Segers, 2014; 2017; Segers & Kockaert, 2016). In our study, the reliability of both methods is compared empirically by scoring the same set of translations (n > 100) according to each method.

1. Introduction

Throughout the years, a substantial number of papers in Translation Studies have dealt with the issue of how to assess translation quality. According to some scholars, it is one of the most intensely debated topics in translation scholarship and practice (Colina, 2009, p. 236). Not only is there controversy surrounding the concept of translation quality itself (Larose, 1998), but the method of assessing quality is also a bone of contention (Waddington, 2001; Williams, 2001). In the course of the past decade, scholars in Translation Studies have repeatedly voiced the need for more empirical research on translation assessment in order to provide a sound methodological basis for the assessment practices that would contribute to the justification of test scores in both educational and professional contexts (Anckaert et al., 2008; Han, 2016; Waddington, 2004).

In this contribution on translation assessment, we investigate two methods that have been put forward to assess translation competence. We define translation competence as the underlying set of knowledge and skills that is put into operation when translating a source text (ST) into a target text (TT). Although translation competence assessment and translation quality assessment are two terms that are not to be conflated – with the latter referring to the assessment of product that can, for example, be the result of a computer translation – both terms are often used interchangeably in Translation Studies (Al-Quinai, 2000; Colina, 2002; Secară, 2005; Stejkal, 2006). This is to be explained by the equation that is automatically presupposed between a translation product and translation competence. 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 of (foreign) language assessment, where the many components that contribute to language proficiency also constitute a "black box" that can be accessed only through eliciting language performance.

Because of the assessments' reliance on performance, the notion of generalization becomes central in examinations as well as selection procedures: when a translation student is awarded a degree or when a candidate is selected for a translator position, the assumption is that he or she is not only capable of translating one (or a few) particular text(s), but that the candidate is competent to translate other texts. The same holds for (foreign-)language students: the certification that is earned on the basis of language tests is taken to imply that the student will be able to demonstrate foreign-language competence in non-test situations.

However, making inferences about individuals' competences on the basis of one (or a few) product(s) or performance(s) is hazardous. Such generalizations are accompanied by measurement error, as is the case in all scientific domains. It is therefore important to document the confidence that can be placed in these generalizations by looking into the reliability of the test scores. However, questions about the reliability of test procedures or assessment methods have remained largely unanswered as a result of the fact that the teaching and testing of translation competence has largely been in the hands of practitioners and translation scholars who have not received any training in psychometrics. During the past decade, we have witnessed a call for empirically driven translation assessment in Translation Studies: scholars have stressed the importance of presenting evidence of test quality when new methods or tests are put forward as measures of translation competence (Anckaert et al., 2008; Conde Ruano, 2005; Eyckmans et al., 2009; Han, 2016; Waddington, 2004). With this article, we hope to contribute to this line of research.

It is important to emphasize that our study targets a summative evaluation context in which the competence level of translators is being assessed and most often also certified (Stejkal, 2006, p. 13). The assessment procedure in summative evaluation can take place either at the end of a training programme or without any reference to a training programme. This is markedly different from formative evaluation. According to Prégent (1990, p. 53), formative evaluation encompasses the attribution of a score alongside constructive feedback in order to raise trainees' awareness of the development of their translation competence. Because of the different consequences of both evaluation procedures – with summative evaluation leading to either a degree or a job, which is not the case in formative evaluation – summative evaluation is considered a high-stake situation in which the need for reliable assessment is considered more compelling.

A short literature review of fairly recent evolutions in translation assessment is presented in the next section. This is followed by a description of the two assessment methods (section 3) that will form the focus of the comparison in the empirical study in section 4. However, since this article centres on a subject that has so scarcely been addressed in Translation Studies, namely reliability, it is important to clarify this concept right from the start.

2. The concept of reliability

Reliability finds its origin in measurement theory, a branch of applied mathematics that deals with the association of numbers with phenomena. Its concerns include an investigation of the kinds of thing that can be measured, how different measures relate to each other, and the problem of error in the measurement processes (Baker, 2001). Apart from the obvious applications of measurement theory in the exact sciences, it has been extensively used in the humanities and it has a long tradition in Language Testing Theory, where tests have been designed to measure individuals' language ability. Since translation assessment involves a quite similar challenge – finding ways to elicit and assess individuals' ability to translate – the field can draw on the same methodology.

One of the main methodological concerns in all fields of measurement is the analysis of the quality of the measurement. Traditionally, evidence about test quality encompasses two technical properties that serve as an indication of the quality and usefulness of a test: reliability and validity. Both reliability and validity have been adopted in measurement theory as means to free assessment of bias and distortion.

Reliability refers to how dependably or consistently a test measures a characteristic. Unfortunately, determining the reliability of tests that measure skills or knowledge is a tricky affair because there are many sources of chance or random measurement error that may threaten the consistency of the measurement. Not only can an individual's performance on a test be influenced by their temporary psychological or physical state (e.g., different levels of anxiety, stress, fatigue, motivation) and any variation in test conditions (e.g., the time and resources that are allowed), the way a performance is scored can also be subject to variation because of the rater's difficulty in maintaining a consistent scoring criterion or because of the divergent scoring methods of different raters (see Eyckmans et al., 2009 and Waddington, 2004, for examples in the domain of translation assessment). Therefore, the reliability of a test in educational measurement is never absolute. It is expressed as a degree, namely the degree to which test scores can be assumed to be unaffected by measurement error. This degree is indicated by a reliability coefficient. It is denoted by the symbol, "r", and ranges between 0 and 1.

The second property of test quality is validity. Validity evidence indicates that there is a link between the test performance and the competence that the test intended to measure. In other words: it refers to those characteristics the test measures and it gives meaning to test scores. In language assessment for example, oral interviews are seen as valid measures of communicative competence, whereas grammar tests are not. In the case of translation assessment, having candidates translate a text from a source language to a target language has always been considered a valid way of verifying whether the candidate can in fact translate between those languages (and in that particular direction). Since the majority of translation tests consist of the task to translate texts and since this task is univocally considered a valid measure of translation competence, validity discussions in translation assessment literature are scant.

In view of the fact that both methods that are under investigation in this study have the translation of a text as their focus, we can assume that the construct validity of both tests is not in question. However, according to the principles and practices of testing theory, the reliability of a test is a prerequisite to its validity, that is, you cannot draw valid conclusions from a test score unless you have gathered evidence indicating that the test is reliable (Anckaert et al., 2013, p. 80; Bachman, 1990, p. 227). Therefore, the comparison and appraisal of the methods in this study will concentrate on their reliability.

3. Evolution in translation assessment

As a result of the fact that translation assessment has long been informed by practice rather than by empirical–quantitative research, its methodology has been based on common sense rather than academic rigour. Furthermore, the epistemological gap between the fields of Language Testing and Translation Studies (Anckaert et al., 2013; Eyckmans et al., 2016; Eyckmans, Segers & Anckaert, 2012) has caused the proposal by one of the founders of the classical Language Testing discipline, Robert Lado (1961), namely, to incorporate translation assessment into Language Testing Theory, to be ignored by the proponents of Translation Studies. In fact, a more scientific approach to the assessment of translation competence did not gain ground before the 1980s, with Juliane House's work (1981) being regarded as seminal in the field of translation assessment (Gile, 2005, p. 254; Lee-Jahnke, 2001). Gradually, analytical methods of assessment came to replace the holistic and intuitive approaches that were generally used to appraise translations. This change in methodology was motivated by the need to objectify the evaluation process.

A bird's-eye view of the Translation Studies literature reveals that researchers who are aware of the fact that measuring translation quality is a subjective process because it relies on human judgement, propose to base the assessment of translation quality on analytical grids which they believe to represent objective evaluation criteria (Al-Qinai, 2000; Horton, 1998; House, 1981; Waddington, 2001). They are supported in this by scholars who are convinced that there is no universally accepted evaluation model in the translation world (Pym, 1992; Sager, 1989; Secară, 2005).

These analytical grids reflect the criteria the grader sees as essential for determining the quality of a translation. They traditionally consist of a near-exhaustive taxonomy of different kinds of mistake of grammar, text cohesion, word choice, etc., combined with a relative weight that is attributed to these categories. Although empirical research has shown that the analytical approach falls short of adequately reducing the subjectivity of the evaluation – mainly because of disagreement on the categorization and weighting of the translation mistakes (Anckaert et al., 2008, p. 55; Anckaert et al., 2013, p. 83; Eyckmans et al., 2009, p. 74) – the use of analytical grids is nowadays widespread in the field of translation assessment (Martínez, 2014; Secară, 2005).

In an attempt to free evaluation of construct-irrelevant variables that are inherent to both holistic and analytical scoring, scholars have taken up new positions in the past decade. Some have decided to abandon the quest for "objective" assessment, investigating instead issues such as inter- and intra-rater reliability, and the construct validity of translation tests (Anckaert et al., 2008; Conde Ruano, 2005; PACTE, 2000; Waddington, 2001). Others have decided to transfer the methodology of educational measurement, and the insights of Language Testing Theory in particular, to the domain of translation assessment by developing norm-referenced instead of criterion-referenced translation tests (Anckaert et al., 2008; Eyckmans et al., 2009).

The norm-referenced approach in translation assessment consists mainly of transferring the well-known "item" concept of standard Language Testing Theory and practice to the field of Translation Studies and translation assessment. The first – and as far as we know the only – attempt to determine translation competence on the basis of a sample-based (i.e., norm-referenced) methodology is the Calibration of Dichotomous Items (CDI) method (Anckaert et al., 2008; Eyckmans et al., 2009). More recently, the Preselected Items Evaluation method was introduced (PIE method; Kockaert & Segers, 2014; 2017; Segers & Kockaert, 2016). This method was developed as a simplified version of the CDI and retains the idea of assessing translation competence on the basis of items. However, in this method, the first and second stages are (called) criterion-referenced instead of norm-referenced (Segers & Kockaert, 2016, p. 70), which is markedly different from the CDI method. Both methods are described in the next section.

4. The CDI method and the PIE method

The written accounts of both the CDI method and the PIE method were perused (Anckaert et al., 2008; Eyckmans et al., 2009; Eyckmans et al., 2012; Kockaert & Segers, 2014; 2017; Segers & Kockaert, 2016) with a view to comparing their methodology and applying them in the empirical study that is reported in section 5.

4.1. General description

In the CDI method, translations are scored on the basis of test-takers' performance on a particular set of translated segments. These segments are determined in a pre-test procedure that is at the root of the method and which involves translating an entire ST. The segments that ensue from the CDI procedure are called "calibrated items" or "items" for short, because the methodology to determine them is drawn from Language Testing Theory and refers to how items are conceptualized in this field, namely, as elements that measure the intended construct in an unbiased way and meet the established principles of language test construction (Bachman, 1990).

The CDI methodology falls under a norm-referenced approach because it is aimed at detecting differences in translation competence among the test-takers. The method hinges on the principle that every element of the text that contributes to the measurement of differences in translation ability between test-takers acquires the status of an item. However, those differences are not determined on the basis of a series of pre-established criteria (which characterizes a criterion-referenced approach); instead, they are determined on the basis of a pre-test procedure in which the performance of a sufficiently large and representative sample of translator trainees is analysed to see which text segments demonstrate discriminating power, that is, which test segments are translated correctly by test-takers who demonstrate good translation competence and incorrectly by test-takers who are not competent translators.

The notion of sampling is central to creating reliable tests in language testing, and it is equally important in the CDI method. Generally, a sample is defined as a group of participants who have been selected from an entire population of possible participants and whose characteristics represent the characteristics of the entire population. The representative nature of the sample is considered important in scientific research because scientists want to extend their findings to a larger group of people, not just those in the study. In the case of the CDI method or the PIE method it means that the sample of translator trainees in the pre-test procedure needs to represent the population of trainees to which the test will apply. When one aims to measure translation competence, translator trainees of all developmental levels need to be included.

The CDI method is based on a "dichotomous" approach, which means that it precludes the weighing of mistakes (or bonuses). It adheres to an assessment policy in which translated segments are either acceptable or not. It has been emphasized that this does not imply that there is only one appropriate translation for the text segment in question; it merely means that for each translated segment it is agreed which alternatives are acceptable and which are not (Eyckmans et al., 2009). When there is disagreement concerning the acceptability of a translation alternative, the item is to be included in the procedure until the statistical analysis of the pre-test allows it to be discarded on the basis of a low discrimination value.

In the PIE-method translations are also scored on the basis of test-takers' performance on a particular set of translated segments, but these segments are not determined on the basis of the performance of a representative group of translation trainees in a pre-test procedure. Instead, the text segments are preselected by the translation grader before the translation trainees translate the text (Kockaert & Segers, 2014). In a next stage, the text is translated by the translation trainees (without their knowing about the identified preselected items) and the translation solutions of the preselected items are evaluated. For these preselected items, the difficulty index is determined (p-value) and on the basis of the p-value, a discrimination index (d-value) is computed (for a detailed account see section 4.2). Only items that show a difficulty index and discrimination index between or above a threshold value (between .2 and .9 for the p-value and above .3 for the d-value) are retained for the final score calculation.

4.2. Procedures of the CDI method and the PIE method

The scores according to the CDI method are obtained as follows: In the pre-test phase, the text to be translated is administered to a representative sample of translator trainees. Then two experienced translation graders go through all the translations and separately register all the segments that they deem to be unacceptable or doubtful translations of the ST. This results in one cumulative list of the total number of translation "errors"[1] that were made by the total number of translation trainees that make up the representative sample. This list is transferred into an Excel matrix with "1" for each time a particular segment of the ST was correctly translated by a particular student and "0" for each time a translation segment was mistakenly translated by a particular student. On the basis of this Excel matrix, the discriminating power of the translation segments is calculated. This is done in SPPS by requesting the corrected item–total correlations (also called the r_{it} value) for all the text segments. The r_{it} value calculates the correlation between the item and the rest of the scale, without that item being considered as part of the scale; that is, it reflects the amount the item contributes to the test's global reliability. To calculate the test score, only items with good discriminating power (> .3; Ebel, 1979) are considered. The reliability of the final selection of test items is calculated by means of Cronbach's alpha.

The scores according to the PIE method are obtained as follows: in the pre-test phase, ten items[2] are preselected from an ST (Kockaert & Segers, 2014; 2017; Segers & Kockaert, 2016). These ten items are identified by the expert grader(s). Correct and incorrect translations of these items are also determined by the expert graders before the test is administered (Kockaert & Segers, 2014, p. 238).[3] Then the ST is handed to the translation trainees to translate. Their performances are screened for their translations of the ten preselected segments. On the basis of these results, the difficulty index is calculated by dividing the number of correct translation solutions by the total number of translations. The discriminatory power of the preselected items is established by substracting the p-value in the bottom group from the p-value in the top group. For the final score calculation, only items with acceptable difficulty and discrimination indices are retained. The scores are expressed as the number of correctly translated items over the total number of discriminating items.[4]

4.3. A theoretical comparison of both methods

The PIE method was developed as an alternative and successor to the laborious CDI method. Its aim is to cut down the number of items to be tested while retaining the reliability advantage (Kockaert & Segers, 2014, pp. 238–239). The method is currently being incorporated in the EULITA project of the European Legal Interpreters and Translation Organisation (EULITA, 2009) that aims to promote best practices in the training and the professional development of legal interpreting and translation in all its member states. A careful evaluation of this new method is therefore warranted. Before moving on to an empirical comparison of both methods, differences in their underlying theoretical and methodological principles are discussed.

One of the most striking observations when reading about the PIE method is the apparent contradiction between the statement of its developers that the assessment of translation competence and the quality control of translation products need different approaches (Kockaert & Segers, 2014, p. 232), and the fact that the PIE method itself does not encompass the evaluation of an entire translation performance. Rather than assessing the quality of a whole translated text, the evaluation is restricted to the correct translation of a few short text segments (between one and six words). One would expect a more elaborate evaluation procedure for assessing a complex skill such as translation competence (ISO 17100:2015; PACTE, 2000) than for judging the quality of one single translated text (in the case of quality control of a translation product). Instead, the scores that are supposed to reflect the translation competence of the translator trainees are calculated on the basis of a mere 5 or 7 items (Kockaert & Segers, 2014, p. 247; Kockaert & Segers, 2017, p. 159; Segers & Kockaert, 2016, p. 75).

A second observation with regard to both methods is that although the use of dichotomous items is promoted in both the CDI method and the PIE method – most likely because it allows one to compute and control the reliability of the assessment, just as it does in language testing – no reliability coefficients have been presented in any of the accounts of the PIE method. Therefore, the claim that "la méthode PIE a l'avantage d'être fiable" (Kockaert & Segers, 2014, p. 248) remains unsubstantiated. Since the validity of a test is contingent on its reliability, no assurances can be given regarding the validity of the method either.

Thirdly, both methods rely on discrimination indices being calculated. In the CDI method, these are computed by using the corrected item–total correlations (r_{it} value) function in SPSS, whereas the PIE method calculates the d-index of potential test items by means of the rather basic "extreme-group method" that dates from the pre-computer era (Pidgeon & Yates, 1968). The reasons for favouring this method are not explained. This is surprising, since the r_{it} value has the advantage over the extreme-group method that every test-taker's score is used to compute the discrimination coefficient, whereas only 54 per cent of test-takers' results are used in the case of the extreme-group method (i.e., the 27% upper and the 27% lower scores: Wiersma & Jurs, 1990).

Apart from how the d-value is computed, the discrimination indices in the PIE method are determined on the basis of astonishingly small samples (ranging from 12 test-takers in Segers & Kockaert, 2016 to 19 test-takers in Kockaert & Segers, 2014, 2017[5]). As a result, the obtained indices have to be interpreted with great caution. The authors' particular interpretation and application of the norm-referenced principle as pertaining to the performance of only a small group of tested candidates (a homogenous group of 12 master's students in Segers & Kockaert, 2016 and 19 master's students in Kockaert & Segers, 2014, 2017) has as a consequence that the generalisations 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.

Finally, there seems to be a significant difference in how the validity argument is addressed in the two methods. The CDI method is designed to steer clear of the pre-test identification of so-called "rich points", namely specific ST segments that contain prototypical translation problems, that is, the most salient, characteristic and difficult problems in a text (PACTE, 2011, p. 322). Rich points reveal a grader's predetermined criteria for translation competence and because this constitutes a source of variability, they are banned from the method. In the CDI method, every possible segment of any given text can acquire the item status on condition that the psychometric analysis reveals the segment's contribution to the measurement of translation competence. The only translation-quality criterion that is applied in the CDI method as an a priori is whether the translation of any given segment of the ST is acceptable or not. The subjectivity involved in this decision is curtailed at a later stage, that is, when the long list of considered items is reduced to a list of items that have survived the process of calibration. In contrast, the developers of the PIE method put forward a procedure in which stage 1 (preselection of the items in the ST) and stage 2 (determination of correct and incorrect solutions of the preselected items) are "criterion-referenced", a term the authors use to refer to the fact that these stages are based on predetermined criteria (Segers & Kockaert, 2016, p. 70). However, these predetermined criteria are not explained or illustrated. It is clear, though, that the preselection of items in stage 1 is done by the authors, who are translation scholars. We can also infer that they have experience in the grading of translations, although this is not reported. It is safe to assume that the criterion the authors have in mind in this preselection stage is the expertise the translator trainer has acquired in grading translations. This raises the compelling question whether different experts would arrive at a similar preselection of items when considering the same text and the same population of test-takers. This question will be taken up in the empirical study.

The theoretical comparison of both methods reveals that the PIE method is less time-consuming and therefore more user-friendly than the CDI method. However, its psychometric properties have not been evinced. Empirical data are needed to throw light on how scores resulting from the application of the PIE method compare to those obtained by the CDI method so far as test quality, in particular reliability, is concerned.

5. Case study

In this case study, we gather empirical data in order to compare the reliability of the two methods discussed. As has already been explained, both methods have been developed with a view to enhancing the reliability of translation assessment and both centre on the use of "items" in order to arrive at translation test scores. However, the way the item-concept is made operational is markedly different: whereas the CDI method uses the performance of a representative group of test-takers to determine discriminating items, the PIE method determines the discriminating power of items on the basis of both graders' expert knowledge in preselecting text segments and the discrimination indices of small groups of participants that have translated the ST.

In order to ward off the reliability threat that arises from the use of small samples in the PIE method, we apply the PIE method to a large, representative sample of translation trainees. The same set of translations (n = 113) is scored according to the procedures of each method and the reliability of both sets of test scores is computed. Since reliability is directly related to measurement error, a comparison of the reliability of both methods will enable us to determine the amount of faith that is warranted in the test scores and in the measurement of the underlying translation competence of the test-takers.

To counter the overreliance on the expertise of one (or two) experienced grader(s) in the first and second stages of the PIE method – which might bias the results of the case study – it was decided to include several expert graders. For each of these graders' sets of preselected items, scores are calculated according to the PIE method. Of these, the set of test scores that shows the highest reliability (i.e., the test scores of the "best" grader) is compared to the reliability of the test scores obtained by the CDI method.

5.1. Research aim

The central research aim concerns the reliability of the proposed translation assessment methods. Since the PIE method is applied by several expert graders, the possibility of obtaining different results for each grader's set of preselected items cannot be ruled out. This possibility is therefore part of the research questions of this empirical study:

1. Do translator trainers who have expertise in grading translations identify the same text segments as preselected items in the first stage of the PIE method? If not, what is the degree of convergence between the preselected items of the different expert graders?
2. How does the reliability of the test scores obtained through the CDI method compare to the reliability of the test scores obtained through the PIE method?

5.2. Participants

5.2.1. Test-takers

A total of 113 translation performances form the empirical basis of this study. This corpus was made up of the Dutch–French translations of all bachelor-level and master-level students from the four francophone translation colleges in Belgium (Haute Ecole de Bruxelles, Haute Ecole Francisco Ferrer, Haute Ecole Léonard De Vinci and Université de Mons-Hainaut[6]). The distribution of student numbers over the four colleges is shown in Table 1.

Table 1: Distribution of participants over the four francophone translation colleges

	Haute Ecole de Bruxelles	Haute Ecole Francisco Ferrer	Haute Ecole Léonard De Vinci	Université de Mons-Hainaut	Totals
1st bachelor year	7	4	0	11	22
2nd bachelor year	13	1	13	12	39
3rd bachelor year	21	2	15	0	38
Master’s year	8	1	5	0	14
Totals	49	8	33	23	113

By including all four translation colleges and all degrees (22 students from the first bachelor year, 39 from the second bachelor year, 38 from the third bachelor year and 14 students from the master’s year) our sample attempts to reflect the entire population of Dutch–French translation students in Belgium. The corpus of translation performances that was used for this case study forms part of a larger corpus of translated texts that was assembled in a previous research design (Eyckmans et al., 2009). The 113 translated texts that were used for the current study are the result of a translation assignment of a short text of 346 words from Dutch into French (see Appendix 1). The text can be characterized as a non-specialized journalistic text about the advertising world. The length, type and level of difficulty of the text conforms to the kind of text material that is used in Belgian students’ translation courses and exams.

5.2.2. Expert graders

In the accounts of the PIE method, the preliminary selection of items was done by the authors who are experienced translator trainers. In order to account for this element of expertise in the application of the PIE method, it was decided to include and compare the item selection of different expert graders in this case study. A total of 20 colleagues from translation colleges and translation services[7] were approached to participate in the study. They had been selected on the basis of their extensive experience in assessing Dutch–French translations. They were sent the Dutch journalistic text and they were invited to identify 10 ST segments for a translation examination.

In the instruction, the term "preselected item" from the PIE method was not used as the participants could not be expected to be familiar with the method or its terminology. Instead, they were asked to select ST segments that a competent translator would be able to translate correctly whereas a translator of intermediate competence would fail to do so (see the instructions for the expert graders in Appendix 1). Seven colleagues (4 male and 3 female) responded positively to the request to take part in the study and also filled in a short questionnaire about their experience in translation assessment (see Appendix 2). Their names have been anonymized in the tables and appendices. Five of the expert graders had French as a mother tongue and two of them were Dutch native speakers. Two of them were professional translators, two were translation trainers and the remaining three combined both professions. Coincidentally, all the francophone translation colleges were represented in this modest collection of expert graders. All of the graders had a minimum of 3 years' grading experience and five of them had more than 10 years' experience in grading translations. Of these five, four individuals claimed that they assessed more than 100 pages a year.

5.3. Procedure

For establishing the scores on the basis of both methods, the reported procedures and stages of the methods were followed, the results of which are described below.

5.4. Results

5.4.1. CDI method

The cumulative list of translation errors that were made by the 113 participants amounted to 170. This number includes those instances in which there was disagreement between the two graders. In each case of disagreement, the translation segment under discussion was included in the list. After transferring the list of 170 translation segments into an Excel matrix, the corrected item–total correlations were calculated by means of the "r_{it} value" function in SPSS in order to distinguish items with good discriminating power from items with low discriminating power. A total of 77 items reached the required threshold (r_{it} value > .3). These are the items that are called "calibrated" and that have been retained in order to calculate the test score and the reliability of the test. For the purposes of this article, standardized scores need not be calculated, therefore the mean and standard deviation of the raw scores are listed in Table 2. Scores range between 3 and 75 on a total of 77. The reliability of the test (containing 77 items) was calculated by means of Cronbach's alpha and amounted to .958.

Table 2: Descriptive statistics of the set of test scores obtained with the CDI method

	N	Range	Min.	Max.	Mean	SD
Score	113	72.00	3.00	75.00	59.39	14.96

5.4.2. PIE method

Each of the seven experts identified 10 items in the ST (cf. Kockaert & Segers, 2014, 2017; Segers & Kockaert, 2016). However, their identifications show little overlap: of the cumulative list of 38 items that had been identified by the seven expert graders, only 14 were selected by more than one grader (see Appendix 3). Only five ST segments were selected by more than half of the graders. These ST segments are "Dat publiek wordt geleverd door de media", "miljardenstroom van reclamegeld", "staat slechts aan de oevers daarvan", "hengelend naar een opdracht", en "waarin het grote reclamegeld omgaat" (for the full ST see Appendix 1). There is not one item that all the expert graders agreed should figure in the selection. Therefore, the PIE method was applied separately for each of these graders' selections of text segments.

The performance of the translator trainees on each of these sets of 10 preselected items was analysed and calculated. On the basis of the results obtained for all seven datasets, the discriminatory power of the preselected items was established. Instead of using the extreme-group method to calculate the discrimination indices (as in Kockaert & Segers, 2014, 2017; Segers & Kockaert, 2016), the more accurate corrected item–total correlations (r_{it} value) were calculated by means of SPSS. Only items with a discrimination coefficient above .30 were retained for the definitive score calculation (Kockaert & Segers, 2014, p. 246). Table 3 shows that the number of preselected items identified by the seven expert graders is reduced considerably when the PIE-procedure is applied: the number of items with acceptable discriminating power ranges from 1 (in the case of grader A and D) to 4 (in the case of grader C).

Table 3: Number of items with sufficient discriminatory power for each expert grader's collection of preselected items. A reliability index has been calculated for the collection that exceeds 3 items

Expert grader	Number of preselected items	Number of items with r_{it} value > .3	Cronbach's alpha on the basis of definitive set of test items
A	10	1	–
B	10	2	–
C	10	4	.558
D	10	1	–
E	10	3	–
F	10	3	–
G	10	3	–

Since so few preselected items remain after the PIE methodology has been applied, the calculation of a test score becomes questionable. Most test developers (or trainers) would feel uncomfortable with basing scores on so few items. This is because the threat of measurement error increases as the test length or the number of items in a test decreases. Indeed, from the point of view of psychometrics the question is whether very few questions or items provide enough measurement information to allow conclusions from the score obtained (e.g., was this translator trainee able to translate the given source test?). In an attempt to be faithful to the PIE method as it is presented in the literature, scores were calculated for the item set of grader C. As can be seen in Table 4, these scores range from 0 to 4.

Table 4: Descriptive statistics of the set of test scores obtained with the PIE method (grader C)

	N	Range	Min.	Max.	Mean	SD
Score	113	4.00	0.00	4.00	2.71	1.08

For the sake of comparing the reliability of both methods that was promised at the start of this study, Cronbach's alpha was calculated for the test scores obtained by grader C (who has the "largest" set of discriminating items, namely 4). The reliability of that particular set of test scores reaches a value of .558.

5.3. Discussion

With reference to the first research question of this study, we observe that there is a poor overlap between the expert graders' selection of rich points. This lack of convergence casts a questionable light on the use of expert graders for preselecting items in the PIE method since the ensuing methodology (i.e., the selection of items to be scored on) is completely dependent on this preselection. If experienced graders choose different rich points, the resulting selection of test items will also be different. This means that the PIE method does not fulfil the objectivity criterion that it professes: the identification of actual test items is constrained by an expert grader's subjectivity in choosing rich points.

The second research question concerns the reliability of the test scores obtained in both methods. When the CDI procedure was applied to assess the translation performances under study, 170 translation errors were identified and 77 items reached the required threshold level to discriminate. With a Cronbach's alpha of .958, the reliability of this set of test scores is high. This is not surprising, since a reliability coefficient is a direct consequence of the selection of items on the basis of their discriminating power. When the PIE method is applied to the same translation performances, the assembled data do not allow a sufficient number of discriminating items to be compiled on which to base a test score.

With so few remaining items, the hope of attaining acceptable test reliability becomes a mirage. It is simply not possible to distinguish reliably between the translation competence of test-takers on the basis of a handful of items. Nevertheless, the reliability was calculated for the test results that were obtained with the PIE method for expert grader C, which amounted to .558. This is a very low reliability coefficient. In applied settings where decisions are made with respect to test scores, a reliability level of .90 is advised as the minimum and .95 is considered the desirable standard (Nunnally, 1978, pp. 245–246). The comparison with the CDI method illustrates that this lack of reliability in the PIE method is not the result of insufficient sample size in this case. With 113 test-takers who are drawn from different levels of translation training programmes the sample is representative enough. The problem lies with the inadequate number of preselected items and their low discriminating power.

5.4. Conclusion

The comparison of the CDI method and the PIE method on the basis of empirical data revealed the inadequacies of the latter. The PIE method did not allow discrimination between the translation performances of the test-takers: the number of items to base a test score on was too low and the reliability of the test scores was below par. In contrast, the CDI method permitted a distinction between the test-takers' translation performances in a reliable way. The difference in the time and effort that are required to apply both methods is apparent. It seems that when it comes to measuring translation competence, short-cuts such as those proposed in the PIE method come at the cost of reliability (and thus validity).

Both the empirical data and the analysis of the underlying theoretical and methodological premises of both methods illustrate the epistemological challenge that Translation Studies faces in transitioning from criterion-referenced to norm-referenced evaluation practices. It also points to the importance of applying norm-referenced principles accurately when developing methods for assessment. With reference to this, translation assessment scholars who want to venture into norm-referenced assessment approaches should heed the following concerns.

First of all, it is important to realize that discrimination indices are statistics, which means that they are measures drawn from a sample. The interpretation we place on them should be informed by our knowledge of the sample. If the sample that is used in the pre-test procedure is not sufficiently large or not representative (cf. the studies reported in the articles on the PIE method), then the statistics that are generated on the basis of it (discrimination indices, for example) cannot be related sensibly to the measured competence.

Secondly, in the case of summative test construction, attempts should be made to obtain large samples of test-takers in order to decrease the standard error of measure of the item's characteristics, as has been professed by many scholars (Bonett, 2002; Carmines & Zeller, 1979; Conrad, 1948; Henryson, 1971; Nunnally, 1978).

One of the many advantages of an interdisciplinary research domain is that it can draw on the knowledge and insights of research fields that fuel the discipline. In the case of translation-quality assessment, Language Testing Theory and methodology can serve Translation Studies well because of what they can offer in terms of control over the design, the validation and the evaluation of translation tests. In this respect, the most promising prospect that norm-referenced testing has in store for the evaluation of translation competence is the fact that the different phases of the procedure ensure that weaknesses in the test or the test use will be exposed (and can be remedied). For example, if the chosen text does not allow discrimination between the competence of the translator trainees because it is too easy or too difficult, this will be brought to light by a low or a moderate reliability.

Finally, a remark concerning the lengthy debate about objectivity and subjectivity might be conducive to the further development of translation assessment. It seems that most discussions centre on the quest for objectivity and on "solving the problem of subjectivity" (Secară, 2005, p. 39), but they should not. As a community of translation scholars and practitioners, we do not need to get rid of subjectivity in translation assessment; quite to the contrary. Also, we are not hindered by the fact that there is no universally accepted evaluation model in the translation world or that there are no absolute standards of translation quality (Sager, 1989). Translation quality and translation competence *should* rely on human judgement (and therefore be subjective). The fact that this subjectivity may throw a wrench in the works when (multiple) translations need to be assessed is more often than not presented as an insurmountable problem in Translation Studies whereas it can be solved, thanks to the methodology of (educational) measurement. We do not need generally accepted objective criteria for evaluating the quality of translations – all the models and translation typologies in the world will not lead to a universally shared consensus for all languages and all texts – what we do need is a sound methodological basis for translation assessment practices in which the subjective process of judging translation quality is embraced and the measurement error that comes with it is calculated, expressed and controlled by means of a reliability coefficient.

References

Al-Qinai, J. (2000). Translation quality assessment: Strategies, parameters and procedures. *Meta, 45*(3), 497–519.

Anckaert, Ph., Eyckmans, J., & Segers, W. (2008). Pour une évaluation normative de la compétence de traduction. *International Journal of Applied Linguistics, 155*, 53–76.

Anckaert, Ph., Eyckmans, J., Justens, D. & Segers, W. (2013). Bon sens, faux-sens, contresens et non-sens sens dessus dessous: Pour une évaluation fidèle et valide de la compétence de traduction. In J.-Y. Le Disez & W. Segers (Eds.), *Le bon sens en traduction* (pp. 79–94). Rennes: Presses universitaires de Rennes.

Eyckmans, J., Anckaert, Ph. & Segers, W. (2009). The perks of norm-referenced translation evaluation. In C. Angelelli & H. Jacobson (Eds.), *Testing and assessment in translation and interpreting* (pp. 73–93). Amsterdam: John Benjamins.

Eyckmans, J., Anckaert, Ph. & Segers, W. (2016). Translation and interpretation skills. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 219–235). Berlin: De Gruyter/Mouton.

Eyckmans, J., Segers, W. & Anckaert, Ph. (2012). Translation assessment methodology and the prospects of European collaboration. In D. Tsagari & I. Csépes (Eds.), *Collaboration in language testing and assessment* (pp. 171–184). Frankfurt am Main: Peter Lang.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker, F. (2001). *The basics of item response theory*. University of Maryland, College Park: ERIC Clearinghouse on Assessment and Evaluation.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Newbury Park, CA: Sage.

Colina, S. (2002). Second language acquisition, language teaching and translation studies. *The Translator, 8*(1), 1–24.

Colina, S. (2009). Further evidence for a functionalist approach to translation quality evaluation. *Target*, *21*, 235–264.

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

Conrad, S. H. (1948). Characteristics and uses of item-analysis data. *Psychological Monographs*, *62*, 1–48.

Ebel, R. L. (1979). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

EULITA. (2009). http://www.eulita.eu/home

Gile, D. (2005). La traduction: La comprendre, l'apprendre. Paris: PUF.

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

Henryson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 153-159). Washington, DC: Council on Education,

Horton, D. (1998). Translation assessment: Notes on the interlingual transfer of an advertising text. *IRAL*, *36*(2), 95–119.

House, J. (1981). A model for translation quality assessment. Tübingen: Gunter Narr.

ISO 17100:2015. Translation Services: Requirements for Translation Services. Technical Committee ISO/TC37, 2015

Kockaert, H. J., & Segers, W. (2014). Evaluation de la traduction: La méthode PIE (Preselected Items Evaluation). *Turjuman. Revue de Traduction et d'Interprétation / Journal of Translation Studies,23*(2), 232–250.

Kockaert H. J., Segers W. (2017). Evaluation of legal translations: PIE method (Preselected Items Evaluation). *Journal of Specialised Translation, 27*, 148–163.

Lado, R. (1961). Language testing: The construction and use of foreign language tests: A teacher's book. London: Longmans.

Larose, R. (1998). Méthodologie de l'évaluation des traductions. *Meta*, *43*(2*)*, 163–186.

Lee-Jahnke, H. (2001). Aspects pédagogiques de l'évaluation en traduction, *Meta*, *46*(2), 258–271.

Martínez, R. (2014). A deeper look into metrics for translation quality assessment (TQA): A case study. *Miscelánea: A Journal of English and American Studies*, *49*, 73–94.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

PACTE. (2000). Acquiring translation competence: Hypotheses and methodological problems in a research project. In A. Beeby, D. Ensinger, & M. Presas (Eds.), *Investigating translation* (pp. 99–106). Amsterdam: John Benjamins.

PACTE. (2011). Results of the validation of the PACTE translation competence model: Translation problems and translation competence. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research: Integrative approaches in translation studies* (pp. 317–343). Amsterdam: John Benjamins.

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

Prégent, R. (1990). *La préparation d'un cours*. Montréal, QC: École polytechnique.

Pym, A. (1992). Translation error analysis and the interface with language teaching. In C. Dollerup & A. Loddegaard (Eds.), *Teaching translation and interpreting. Training, talent and experience: Papers from the first Language International Conference*, Elsinore, Denmark, 31 May–2 June, 1991 (pp. 279–288). Amsterdam: John Benjamins.

Sager, J. C. (1989). Quality and standards: The evaluation of translations. In C. Picken (Ed.), *The translator's handbook* (pp. 91–102). London: ASLIB.

Samuels, P. (2015). Statistical methods: Scale reliability analysis with small samples, Birmingham: Birmingham City University, Centre for Academic Success.

Secară, A. (2005). Translation evaluation: A state of the art survey. Proceedings of the eCoLoRe/MeLLANGE Workshop Leeds (pp. 39–44). Manchester: St. Jerome.

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

Stejkal, J. (2006). Quality assessment in translation. *MultiLingual, 80*(17), 41–44.

Waddington, C. (2001). Different methods of evaluating student translations: The question of validity. *Meta,46*(2), 331–325.

Waddington, C. (2004). Should student translations be assessed holistically or through error analysis? *Lebende Sprachen*, *49*(1), 28–35.

Wiersma, W., & Jurs, S.G. (1990). *Educational measurement and testing* (2nd ed.). Boston, MA: Allyn and Bacon.

Williams, M. (2001). The application of argumentation theory to translation quality assessment. *Meta*, *46*(2*)*, 327–344.

Williams, M. (2009). Translation quality assessment, *Mutatis Mutandis*, *2*(1), 3–23.

Appendix 1

Consignes

Sachant que le texte ci-dessous devra servir de source pour un examen de traduction du néerlandais vers le français, surlignez 10 segments dont vous pensez qu'ils seront mal traduits par des candidats médiocres et bien traduits par ceux qui possèdent une réelle compétence de traduction.

Les segments de texte que vous allez sélectionner, comporteront de 1 à maximum 6 mots.

Numérotez vos segments pour être sûr(e) d'en avoir sélectionné 10 exactement.

Media en amusement trekken alle aandacht naar zich toe

Een van de grootste misvattingen die ik bij de meeste mensen heb vastgesteld over reclame, is dat zij ervan uitgaan dat al het geïnvesteerde reclamegeld van bedrijven terechtkomt in de 'reclamewereld', waarmee zij de wereld van reclamebureaus en reclamemakers bedoelen. Dat is niet zo. Het overgrote deel van de reclame-investeringen van het bedrijfsleven gaat naar de aankoop van ruimte in de media, en komt dus terecht op de bankrekeningen van de mediagroepen met hun tijdschriften, kranten, radiozenders, tv-stations, bioscopen, billboards... Bedrijven zijn immers op zoek naar een publiek om hun producten bekend en geliefd te maken, in de hoop dat publiek ervan te kunnen overtuigen hun producten ten minste eens te proberen. Dat publiek wordt geleverd door de media. De bedrijven kopen dus pagina's of zendtijd, en kunnen zo in contact treden met het publiek van die media. Op die manier ontstaat er een miljardenstroom van reclamegeld (in België meer dan 1,75 miljard euro per jaar), die van de bedrijven naar de media stroomt.

De reclamebureaus staan slechts aan de oevers daarvan. Zij zijn de kleine vissers, hengelend naar een opdracht van de bedrijven die er vooral in bestaat de aangekochte ruimte te vullen met inhoud. De reclamebureaus zorgen dus voor de ontwikkeling van de boodschap van het merk en voor de verpakking van die boodschap. In ruil daarvoor krijgen ze een percentage van de reclame-investeringen: vroeger ging dat om 15%, nu is het meestal minder dan 10%. Steeds meer worden deze percentages afgeschaft en vindt de betaling plaats via een maandelijks honorarium, dat door de bedrijven zwaar onder druk wordt gezet. Zij moeten immers steeds meer betalen voor hun mediaruimte en willen die meerkosten zo veel mogelijk terugverdienen, onder meer via de reclamebureaus. Deze worden verplicht steeds sneller te werken voor steeds minder geld. Dat wil niet zeggen dat de reclamebureaus armoedezaaiers zijn. Veel reclamemakers verdienen goed. Door mee te surfen op de golven van de reclame-investeringen zijn er multinationale en beursgenoteerde reclamenetwerken ontstaan. Maar de échte reclamewereld, waarin het grote reclamegeld omgaat, is eigenlijk de mediawereld.

Appendix 2

Votre profil en tant que sujet d'expérience

1. Vous êtes

❑ une femme
❑ un homme

2. Votre langue A est

❑ le français
❑ le néerlandais
❑ une autre langue

3. Vous êtes

❑ un(e) enseignant(e)
❑ un(e) professionnel(e) de la traduction
❑ les deux

4. Vous êtes diplômé(e) en

❑ traduction - interprétation
❑ philologie germanique
❑ les deux
❑ autre

5. Depuis combien d'années êtes-vous amené(e) à évaluer la qualité de traductions françaises produites à partir de textes sources en néerlandais?

❑ Moins de 3 ans
❑ De 3 à 5 ans
❑ De 6 à 10 ans
❑ Plus de 10 ans

6. A combien estimez-vous le nombre de pages de traductions NL-FR que vous êtes/avez été amené(e) à évaluer en moyenne chaque année?

❑ moins de 10
❑ 10 à 20
❑ 20 à 50
❑ 50 à 100
❑ plus de 100

Appendix 3: Rich points as identified by the seven expert graders

Source-text segment	A	B	C	D	E	F	G	Total
trekken alle aandacht naar zich toe	0	1	0	0	0	1	0	2
Een van de grootste misvattingen	0	0	0	0	0	0	1	1
bij de meeste mensen heb vastgesteld	0	0	0	0	1	0	0	1
is dat zij ervan uitgaan	1	0	0	0	0	0	1	2
Waarmee	0	0	1	0	0	0	0	1
terechtkomt in de 'reclamewereld'	0	1	0	0	0	0	0	1
Dat is niet zo.	1	0	0	0	0	0	0	1
van het bedrijfsleven	0	0	0	0	1	0	0	1
gaat naar de aankoop van ruimte	0	1	0	1	0	0	0	2
met hun tijdschriften	0	0	1	0	1	0	0	2
bekend en geliefd te maken	0	1	0	0	0	1	0	2
ten minste eens te proberen	0	0	0	1	0	0	0	1
naar een publiek om hun producten	0	0	0	0	1	0	0	1
Dat publiek wordt geleverd door de media	1	1	0	1	0	1	1	5
op die manier	0	0	1	0	0	0	0	1
miljardenstroom van reclamegeld	1	0	1	1	1	1	0	5
Stroomt	1	0	0	0	0	0	0	1
Staat slechts aan de oevers daarvan	1	1	1	1	1	1	1	7
zijn de kleine vissers	0	0	0	0	0	1	0	1
hengelend naar een opdracht	1	1	0	0	0	1	1	4
die er vooral in bestaat	0	0	0	0	0	0	1	1
te vullen met inhoud	0	0	1	0	0	0	1	2
De reclamebureau zorgen dus voor	1	0	0	0	0	0	1	2
de verpakking	0	0	0	1	0	0	0	1
ging dat om 15%	0	0	0	1	0	0	0	1
vindt de betaling plaats	0	0	1	0	0	0	0	1
Via	1	0	0	0	0	0	0	1

zwaar onder druk wordt gezet	0	1	0	1	1	0	0	3
moeten immers steeds meer betalen	0	0	0	0	1	0	0	1
Mediaruimte	0	0	0	1	0	0	0	1
steeds minder	0	0	1	0	0	0	0	1
Dat de reclamebureaus armoedezaaiers zijn	0	0	0	0	0	0	1	1
verdienen goed	0	0	1	0	0	0	0	1
Door mee te surfen op	0	1	0	0	1	1	0	3
op de golven van de reclameinvesteringen	0	0	0	0	0	1	0	1
multinationale en beursgenoteerde reclamenetwerken ontstaan	0	0	0	0	1	0	0	1
waarin het grote reclamegeld omgaat	1	1	1	1	0	1	1	6

[1] By "cumulative list of translation errors" we mean a list that includes all translations of a particular ST segment that were considered unacceptable or doubtful by one of the experienced graders.

[2] In both publications the number of preselected items seems to be independent of the number of words of the ST: in Kockaert and Segers (2014, 2017) ten words or word combinations are preselected in a ST of 137 running words; in Segers and Kockaert (2016) ten words or word combinations are preselected in a ST of 339 running words.

[3] "Avant l'épreuve les solutions correctes et incorrectes sont déterminées pour chaque item présélectionné dans le texte source. Les solutions doivent être clairement correctes ou incorrectes. Les évaluateurs éviteront les zones floues" (Kockaert & Segers, 2014, p. 238).

[4] With a maximum score of 7 in Segers and Kockaert (2016, p. 75) and 5 in Kockaert and Segers, (2014, p. 247; 2017, p. 159).

[5] The empirical data that are reported in Kockaert and Segers (2014) and in Kockaert and Segers (2017) seem to be exactly the same (participants, materials, result analysis) but for the sake of completeness we have decided to mention all the publications.

[6] These are the names of the four francophone translation colleges at the time of the data collection. They have since been reorganized into university faculties of the Université Libre de Bruxelles, the Université Saint-Louis, the Université Catholique de Louvain and the Université de Mons-Hainaut.

[7] The term "translation services" refers to government agencies or departments that verify the equivalence of official source documents and their translations.