Evaluation of legal translations: PIE method (Preselected Items Evaluation)

Hendrik J. Kockaert, KU Leuven and University of the Free State, and Winibert Segers, KU Leuven

ABSTRACT

This article is situated in the field of translation evaluation and consists of two parts. The first part is dedicated to the different meanings of the term 'translation evaluation' and the implications for evaluation approaches. The term 'translation evaluation' may refer to the translation product, the translation process, the translation service and the competence of the translator. Product, process, service and competence of the translator require different evaluation approaches. In the second part of the article we illustrate in a case study a method for evaluating the translation product: the PIE method (Preselected Items Evaluation).

KEYWORDS

Translation evaluation, PIE method, legal translation, reliability, validity, p value, d index.

1. Preliminary terminological remarks

Evaluation is giving a rating[1] (e.g. 4/10, 15/20 …) or an evaluative letter (e.g. A = excellent, B = good ... E = very bad) or making a value judgment (e.g. excellent, good … very bad).

The term 'translation evaluation' can refer to the translation product (the target text, the result of the translation process), the translation process (the way the translator goes between the source text, and the target text), the translation service (contact with the client, providing a quote, invoicing, compliance agreements, complaints, etc.), and the competence of the translator.

The product (1), the process (2), the service (3) and the competence of the translator (4) cannot be evaluated in one and the same way. On the basis of the target text, for example, one cannot make a value judgment about the translation process or about the competence of the translator. Product, process, service and competence require different evaluation approaches.

Evaluation must be strictly distinguished from revision and proofreading. Evaluators and revisers often use the comparison-source-text-target-text-method, but the purpose of this comparison is very different for evaluators and revisers.

2. Translation product

The translation product can be evaluated by different methods: the holistic method, the analytical method, the CDI (Calibration of Dichotomous Items) method, and the PIE method (Preselected Items Evaluation).

2.1. Holistic method

The holistic method is a commonly used method in education and in the professional field. The evaluator reads the translation and gives a rating or evaluative letter (e.g. A = excellent, B = good ... E = very bad), and considers the translation as a whole, without analysing the translation in detailed error categories. This value judgment is based on an overall impression. This method is fast, but subjective, and is easily determined by the personal 'taste' of the evaluator. The value judgments of different holistic evaluators on the same translation can vary greatly: what one evaluator considers as a beautiful, creative translation, can be considered as an unacceptable translation by another evaluator (Anckaert et al. 2008; Anckaert et al. 2013; Eyckmans et al. 2009; Eyckmans et al. 2012). In other words, interrater reliability is low. In the margin of the scope of this article, it is interesting to observe that "points-based error focused grading" has been replaced in the University of Helsinki by holistic grading methods (Garant 2009: 7). Garant's main finding is that translation was better assessed with a focus on discourse level holistic evaluation, instead of a detailed point-based, grammar-like, "analytical" evaluation which seems more appropriate for assessment in language proficiency tests. A holistic approach seemed to focus better on a context-sensitive evaluation (e.g. at paragraph level), and seemed to move away from exclusive attention to grammatical errors in translation tests. While these findings are not within the scope of this article, they may however deserve closer attention when we attempt to propose a comprehensive approach to an objective, valid, reliable and practical evaluation method of translation tests in the framework of e.g. certified translation claimed to be essential by McAlester (2000: 231) and recommended by Qualetra[2] and TransCert[3].

2.2. Analytical method (assessment grids)

The analytical method is based on error analysis in assessment grids, and is generally claimed to "[be] more reliable and valid than holistic methods" (Waddington 2001: 316). The analytical evaluator makes use of a grid, a matrix consisting of a number of error types and a number of error levels.

Example of an assessment grid

Error type	Error level		
content			
1 intelligibility	a	b	c
2 coherence	a	b	c

3 other aspects		a	b	c
form				
4 spelling		a	b	c
5 grammar	a	b	c	
6 vocabulary	a	b	c	
7 style		a	b	c
8 other aspects		a	b	c

(a= major error, c = minor error)

Categories 3 and 8 'other aspects' of this grid are open categories. They will be used for errors that cannot be classified as 1, 2, 4, 5, 6 or 7. The number of error types and error levels could be increased (e.g., 30 error types and 10 levels of error), but this increase should be done prudently. An increase could reduce the practical applicability of the grid.

Other examples of assessment grids include the Framework for Standardized Error Marking (American Translators Association), ITR BlackJack, QA Distiller and the SAE J2450 Translation Quality Metric.

The analytical method requires more time than the holistic method, but the translator will have a better understanding of what is right and what is wrong in his translation.

Compared to the holistic method, the analytical method has the disadvantage that the evaluator focussing on small text segments does not necessarily have an overview of the target text as a whole.

The analytical method is no guarantee of objectivity. Different evaluators do not always agree with each other: the same error can be for one evaluator a slight misspelling and for another evaluator a serious grammar mistake (see Anckaert et al. 2008; Anckaert et al. 2013; Eyckmans et al. 2009; Eyckmans et al. 2012).

To solve the problem of subjectivity, the PIE method (Preselected Items Evaluation) was developed.

2.3. PIE method (Preselected Items Evaluation)

The PIE method is an adapted, practical, pragmatic version of the CDI (Calibration of Dichotomous Items) method (Anckaert et al. 2008; Anckaert et al. 2013; Eyckmans et al. 2009; Eyckmans et al. 2012; Kockaert and Segers 2012, 2014). For efficiency and time management reasons, the number of preselected items is limited in the PIE method. Questions can be asked about the ideal number of items (as much as needed, but not more, to allow a reliable and valid evaluation). The CDI method and the PIE method are both calibration methods: the accuracy of the measuring instrument is checked and adjusted. The CDI method and the PIE method are both dichotomous methods: the methods make the distinction between correct and wrong solutions; the methods do not distinguish between levels of error.

With the CDI method, the items to be evaluated are selected on the basis of the calculation of p values and d indices of each item (words, word groups, etc.) in the source text, for which correct and erroneous solutions are determined.

Item difficulty: p value

As is the case with classical item analysis in statistics, CDI includes, and PIE may include, calculations in order to gauge difficulty (prevalence of correct answers) and discrimination (ability to differentiate candidates on the item being measured) for each item in the CDI method, and for a number of preselected items in PIE. We calculate item difficulty values (p values) and item discrimination indices (d index) for the purpose of determining the "minimum number of items needed for a desired level of score reliability or measurement accuracy" (Lei and Wu 2007: 527). Item difficulty is the percentage of candidates who answer the item correctly. The larger the group of candidates answering correctly, the easier the item. The higher the difficulty value or p value, the easier the item. To calculate the item difficulty with the p value, we divide the number of candidates answering the item correctly by the total number of candidates answering item. The proportion for the item is usually denoted as p and is called item difficulty (Crocker and Algina 1986). An item answered correctly by 85% of the examinees would have an item difficulty, or p value, of .85, whereas an item answered correctly by 50% of the examinees would have a lower item difficulty, or p value, of .50 (Matlock-Hetzel 1997: 2).

Item discrimination: d index

After calculating the item difficulty (p values of the items), the discriminating power of the preselected items are calculated (discrimination indices of the preselected items):

> If the test and a single item measure the same thing, one would expect people who do well on the test to answer that item correctly, and those who do poorly to answer the item incorrectly. A good item discriminates between those who do well on the test and those who do poorly. Two indices can be computed to determine the discriminating power of an item, the item discrimination index, D, and discrimination coefficients (Matlock-Hetzel 1997: 5).

In the calculation of the d index, we refer to Wiersma and Jurs (1990), and apply the method of extreme groups: the discriminating power of an item can be measured when we compare the number of candidates with high test scores who answered a particular item correctly with the number of people with low scores who answered the same item correctly. The method of extreme groups calculates the discrimination index with the following parameters: the 27% of the candidates at the top and the 27% at the bottom of the entire score ranking are separated for the analysis. Wiersma and Jurs argue that "27% is used because it has been shown that this value will maximize differences in normal distributions while providing enough cases for analysis" (1990: 145). The discrimination index, d index, is the number of candidates in the top group (i.e. 5 candidates) who answered the item correctly minus the number of candidates in the bottom group (i.e. 7 candidates) who answered the item correctly.

In the CDI method, items which do not respond to the above mentioned docimologically recurrent observations (too high or too low p values, weak discriminating power) will be removed from the translation test and will be replaced by other items so that a critical mass of items can be assured.

The PIE method is characterised by a preselection (before administering the translation test) of items in the source text, on the basis of translation brief relevancy, domain specific and test specific criteria. As is the case with the CDI method, correct and wrong solutions are listed for each preselected item in the source text of the test.

Preselected items may relate to different error types: grammar, spelling, style, vocabulary, etc.

After administering the test, the difficulty of the preselected items will be calculated (p values of the preselected items), and the discrimination power of the preselected items will be calculated (discrimination indices of the preselected items).

Preselected items that do not respond to docimological standards (too high or too low p values, weak discriminating power) may be removed from the translation test and replaced by other items.

Evaluators using the PIE method will have the same value judgment on the translation product. The PIE method ensures objectivity, cross candidate transparency and equality in scoring, hence marking attitude.

With PIE, the validity of the items with very good p values and d indices is strengthened by preselecting items on the basis of translation brief relevance. In this way, PIE based items give a correct image of what the evaluator of a translation product wants to measure (translation brief) and are docimologically justifiable. The difference between CDI and PIE relies in the fact the items in the CDI are selected on the basis of the only docimological dimension, while PIE calculates, optionally, docimological values on translation brief relevant items only. In other words, docimological exactitude complements in a second, and optional, phase, translation brief relevance.

A question which may come up when discussing the docimological impact of a translation test is the following: What do evaluators do with a candidate who proposes an incorrect solution for an item that was not preselected? To answer this question we have to look at the performance of all the candidates who participated in a translation test. If the candidate is the only one of all the candidates who proposes a wrong solution for the not preselected item, this item must not be included in the translation test. But if the item has a good p value and a very good discrimination index (d = 0.40 or higher), this item may be considered for inclusion in the translation test.

Whether we need to consider docimological exactitude and/or translation brief relevance will depend on the criteria of each test: for e.g. entrance tests in international institutions or university programmes, we will need to justify docimologically the selection of the items to evaluate for the purpose of cross candidate objectivity, transparency and equality. When the evaluator needs to recruit e.g. legal translators in the area of criminal proceedings, the selection of items will rely much more on translation brief criteria, such as terminological, phraseological and legal items related to source and target items which are crucial for grasping possible differences between source and target specific features in criminal proceedings. In this context, the advantage of PIE relies on the possibility to adopt PIE without p value and d index calculation (PIE Light), or to adopt PIE complemented in a second phase with p value and d index calculation. What PIE delivers in both evaluation paths, is the inclusion of translation specific requirements.

Another issue is which solutions are correct and which are not. In this context, the PIE method practises a dynamic approach in that next to the predetermined correct solutions by evaluators, unexpected correct solutions are included in the lists (TM-like data base). This approach accords well with Bowker (2001: 346) who supports the search for unexpected correct solutions in online corpora of authentic domain-specific documents for the purpose of giving objective corpus-based evidence, hence documented feedback to translation students.

Table 1 illustrates the differences between the evaluation methods discussed in this article.

Table 1: Evaluation methods: overall comparison

	Holistic	Analytical	CDI	PIE
Number of items	Exhaustive	Exhaustive	Docimologically relevant items	*Translation brief* relevant items
Evaluation	Global	Grids/Criteria	Grids/Criteria	Grids/Criteria

Dichotomous	-	-	√	√
Calibration	-	-	√	√
Acceptance of alternatives	expected/ unexpected	expected/ unexpected	expected/ unexpected	expected/ unexpected
ISO 17100 compatible	√	√	√	√
Interrater reliability	-	-	√	√
Criterion referenced				+
Norm referenced			+	

Table 2 shows that PIE evaluates items which are different from the items selected according to the p values and d indices, which are calculated on the basis of docimological relevance, which does not necessarily coincide with translation brief relevant criteria.

PIE is applied without the optional calculation of p values and d indices (column 1 of Table 2) and CDI is applied after the calculation of p values and d indices (column 2 of Table 2). Column 3 displays the preselected items that were considered crucial by reviewers in the translation brief. When PIE + CDI is adopted, the list of items excludes docimologically non justifiable items from the list of translation brief relevant items (strikethrough and red items in column 4 of Table 2).

Table 2: Selection of items on the basis of CDI versus PIE relevance criteria (See case study below: 'Jugement correctionnel. Demande de mise en liberté').

PIE	**CDI** (p values and d indices)	**PIE** (translation brief relevance: legal translation of criminal proceedings)	**PIE** (translation brief relevance: legal translation of criminal proceedings) **+ CDI**
Après			
Plus			
De		de	
douze			
heures			
De			
délibéré	Délibéré	délibéré	délibéré
,			
La	La		~~la~~
cour	cour	cour	cour
d'	d'	d'	d'
assises	assises	assises	assises

De		de	~~de~~
Paris			
a rendu	a rendu	a rendu	a rendu
Ce			
vendredi			
Son	son		
verdict	verdict	verdict	verdict
dans			
Le			
procès		procès	~~procès~~
De			
La	La		~~la~~
prise	prise	prise	prise
d’	d’	d’	d’
otages	otages	otages	otages
Du			
Ponant			
Le	le	le	le
4	4	4	4
avril	avril		~~avril~~
2008	2008		~~2008~~

Table 3: CDI versus PIE

CDI	PIE
Same value judgment among evaluators	Same value judgment among evaluators
Reliable and less subjective	Reliable and less subjective
Reinforces its potential as assessment method for a more reliable and valid certification of translation competence	Reinforces its potential as assessment method for a more reliable and valid certification of translation competence
Items selected on the basis of docimological criteria	Items selected on the basis of translation brief criteria Option: Translation brief relevant items reselected on the basis of docimological

3. Translation process

The translation process can be evaluated by means of think aloud protocols, eye tracking (tracking the eyes of the translator) and key logging (recording the keystrokes).

4. Translation service

On the basis of a service standard (e.g. ISO 17100 *Translation Services -- Requirements for translation services*) we can evaluate the service. An accreditation body needs to audit the translation service provider to see whether the services rendered by the Translation Service Provider (TSP) are in accordance with the standard.

5. Competence of the translator

In order to evaluate the competence of the translator a competence measurement must be performed. The competence of the translator can be divided into five sub-competences (ISO 17100):

- translation competence
- linguistic competence (source language and target language)
- cultural competence (source culture and target culture)
- research competence
- technical competence

For each of the five sub-competences valid and reliable instruments must be developed[4]. A valid instrument is an instrument that measures what it purports to measure. If such an instrument claims to measure linguistic competence, it is not fit for measuring cultural competence. A reliable instrument is an objective, evaluator-independent tool. The measurement result is not influenced by the evaluator.

Examples of translation competence models

- EMT Competences for professional translators, experts in multilingual and multimedia communication (EMT Expert Group qtd. in Chodkiewicz 2012)
- ISO 17100. Translation Services - Requirements for translation services
- PACTE Process in the Acquisition of Translation Competence and Evaluation (PACTE 2005, 2009)

6. Application of the PIE method – Case Study

Source text: 'Jugement correctionnel. Demande de mise en liberté'. Cour d'Appel de Paris. Tribunal de Grande Instance de Bobigny. Jugement du : 21/08/2012. 17ème chambre correctionnelle. N° minute : 1681/12. P. 2.
Target text: Dutch
Date of the translation test: 03/05/2013 (10.30-12.00 h.)
Number of words: 118
Maximum testing time allowed: 90 minutes
Number of candidates: 19 (students in the MA Translation and Interpretation, KU Leuven, Belgium; native speakers of Dutch)
Evaluation method: PIE method complemented with p values and d indices for docimological justification (CDI)
Number of preselected items: 10
Number of preselected words: 30 (25% of the total number of words)

Students were allowed to use dictionaries and internet resources.

6.1. French source text

JUGEMENT CORRECTIONNEL
DEMANDE DE MISE EN LIBERTÉ
DEBATS

Avant l'audition de XXX, le président a constaté que celui-ci ne parlait pas suffisamment la langue française ;
Il a désigné YYY, interprète inscrit sur la liste du tribunal ; l'interprète a ensuite prêté son ministère chaque fois qu'il a été utile.
A l'appel de la cause, le président a donné connaissance de l'acte qui a saisi le tribunal et constaté la présence et l'identité de XXX, dont il a reçu les déclarations.
Maître ZZZ, conseil du prévenu, a été entendu en sa plaidoirie.
Le ministère public a été entendu en ses réquisitions.
Le prévenu a eu la parole en dernier.
Le greffier a tenu note du déroulement des débats.

6.2. Preselected items

Ten items were preselected in the French source text on the basis of domain specific relevance, and translation brief criteria (Translate the text in Dutch; use the same register and style, appropriate for a legal audience). The preselected items are underlined below and appear in Table 4.

JUGEMENT CORRECTIONNEL
DEMANDE DE MISE EN LIBERTÉ
DEBATS

Avant l'audition de [PI 1] XXX, le président a constaté que celui-ci ne parlait pas suffisamment la langue française ;
Il a désigné YYY, interprète inscrit sur la liste du tribunal ; l'interprète a ensuite prêté son ministère [PI 2] chaque fois qu'il a été utile.
A l'appel de la cause [PI 3], le président a donné connaissance de [PI 4] l'acte qui a saisi le tribunal [PI 5] et constaté la présence et l'identité de XXX, dont [PI 6] il a reçu les déclarations [PI 7].
Maître ZZZ, conseil [PI 8] du prévenu, a été entendu [PI 9] en sa plaidoirie.
Le ministère public a été entendu en ses [PI 10] réquisitions.
Le prévenu a eu la parole en dernier.
Le greffier a tenu note du déroulement des débats.

Table 4: Preselected items

Number of preselected item	Preselected item
1	l'audition de (the hearing of)
2	a … prêté son ministère (exercised his duties)
3	A l'appel de la cause (at the appeal of the case)
4	a donné connaissance de (has given knowledge of)
5	qui a saisi le tribunal (who brought a claim to the Court)
6	dont (of which)
7	les déclarations (the statements)

8	conseil (counsel)
9	a été entendu (was heard)
10	ses (his)

6.3. P values of the preselected items

These values are calculated according to the method described above. The p value corresponds to the number of correct solutions divided by the total number of candidates.

Table 5: P values of the preselected items

Nº. preselected item	**P value**	**Candidate Numbers**	
		Correct solution	**Incorrect solution**
1	l'audition de P value (16/19) = 0.84	1 2 4 5 6 7 8 9 10 12 13 14 15 17 18 19	3 11 16
2	a … prêté son ministère P value (17/19) = 0.89	1 2 3 4 5 6 7 8 10 11 12 13 15 16 17 18 19	9 14
3	A l'appel de la cause P value (12/19) = 0.63	1 3 4 6 9 10 11 13 16 17 18 19	2 5 7 8 12 14 15
4	a donné connaissance de P value (15/19) = 0.79	1 2 3 4 7 8 10 11 12 13 14 16 17 18 19	5 6 9 15
5	qui a saisi le tribunal P value (10/19) = 0.53	2 4 6 7 12 13 14 16 17 18	1 3 5 8 9 10 11 15 19
6	dont P value (15/19) = 0.79	1 3 4 6 8 9 10 11 12 13 14 15 16 17 19	2 5 7 18
7	les déclarations P value (14/19) = 0.74	1 2 3 4 6 7 9 11 12 14 15 16 18 19	5 8 10 13 17
8	conseil P value (14/19) = 0.74	1 2 3 4 7 8 9 11 12 13 14 16 17 19	5 6 10 15 18
9	a été entendu P value (15/19) = 0.79	1 2 4 6 7 8 10 11 12 13 15 16 17 18 19	3 5 9 14
10	ses P value (16/19) = 0.84	2 3 4 6 7 8 9 10 11 12 13 14 16 17 18 19	1 5 15

P values should be higher than 0.20 and lower than 0.90 (Sabri 2013). The p values of the ten preselected items range between 0.53 and 0.89. Preselected item 5 ('qui a saisi le tribunal') is the most difficult item (p = 0.53). Preselected item 2 ('a … prêté son ministère') is the easiest item (p = 0.89).

6.4. Discrimination index (d index) of the preselected items

To measure the d index of the ten preselected items, we use the above explained method of extreme groups.

Table 6: Extreme Group Method: Top and Bottom Groups

Candidate	Score/10	Top Group Nº.	Score/10	Bottom Nº.	Score/10
1	8	4	10	5	2
2	8	12	9	15	5
3	7	13	9	9	6
4	10	16	9	3	7
5	2	17	9	8	7
6	8	19	9	10	7
7	8	1	8	14	7
8	7	2	8	1	8
9	6	6	8	2	8
10	7	7	8	6	8
11	8	11	8	7	8
12	9	18	8	11	8
13	9	3	7	18	8
14	7	8	7	12	9
15	5	10	7	13	9
16	9	14	7	16	9
17	9	9	6	17	9
18	8	15	5	19	9
19	9	5	2	4	10

Table 7: D indices regardless of docimological impact

Item	p value Top Group	p value Bottom Group	d index[5]
1	0.80	0.86	-0.06
2	1.00	0.71	0.29
3	0.80	0.43	0.37
4	1.00	0.57	0.43
5	1.00	0.14	0.86
6	1.00	0.86	0.14
7	0.60	0.57	0.03
8	1.00	0.57	0.43
9	1.00	0.43	0.57
10	1.00	0.71	0.29

On the basis of the d index calculations in Table 7, we conclude that the d index of items 1, 2, 6, 7 and 10 is docimologically unjustified because of its value inferior to < 0.30.

6.5. Accepted items after the docimologically justified calculation of the p values and the d indices of the ten preselected items

Table 8: Items with docimologically justified d indices

Item	p value	d index*
3	0.63	0.37
4	0.79	0.43
5	0.53	0.86
8	0.74	0.43
9	0.79	0.57

On the basis of the five accepted items we can recalculate the scores of the nineteen candidates as is shown in the table below.

Table 9: Recalculated scores on the basis of docimologically justified d indices

Candidate	Score /10	Score /5
1	8	4
2	8	4
3	7	3
4	10	5
5	2	0
6	8	3
7	8	4
8	7	3
9	6	2
10	7	3
11	8	4
12	9	4
13	9	5
14	7	3
15	5	1
16	9	5
17	9	5
18	8	4
19	9	4

The consequence of this recalculation is the most severe for candidate 15, who goes from 5/10 to 1/5. Candidate 15 has correctly translated preselected items 1, 2, 6, 7 and 9. But four of these five items (items 1, 2, 6, and 7) are not accepted after calculation of the p values and the d indices.

7. Conclusion

The PIE method offers the advantage of reliability in the context of translation evaluation: each candidate is evaluated on the same items, which have been selected on the basis of the translation brief relevance. Optionally, the preselected items may be tested on their docimological strength on the basis of p values and d indices. This fully-fledged evaluation method complements reliability with test validity: the correlation between scores obtained on the translation test and evaluated with PIE, and scores obtained and evaluated with, for example, analytical grid method, or the CDI method.

Bibliography

- **Anckaert, Philippe, Eyckmans, June and Winibert Segers** (2008). "Pour une évaluation normative de la compétence de traduction. " *ITL Review of Applied Linguistics* 41/155, 53-76.

- **Anckaert, Philippe, Eyckmans, June, Justens, Daniel and Winibert Segers** (2013). "Bon sens, faux sens, contresens et non-sens sens dessus dessous: pour une évaluation fidèle et valide de la compétence de traduction." Jean-Yves Le Disez and Winibert Segers (eds) (2013). *Le bon sens en traduction*. Rennes: Presses Universitaires de Rennes, 79-93.

- **Bowker, Lynne** (2001). "Towards a Methodology for a Corpus-Based Approach to Translation Evaluation." *Meta: Translators' Journal* 46(2), 345-364.

- **Chodkiewicz, Marta** (2012)."The EMT framework of reference for competences applied to translation: perceptions by professional and student translators." *The Journal of Specialised Translation* 17, 37-54.

- **Eyckmans, June, Anckaert, Philippe and Winibert Segers** (2009). "The perks of norm-referenced translation evaluation." Claudia Angelelli and Holly Jacobson (eds) (2009). *Testing and assessment in translation and interpreting studies*. Amsterdam: John Benjamins, 73-93.

- **Eyckmans, June, Segers, Winibert and Philippe Anckaert** (2012). "Translation assessment methodology and the prospects of European collaboration." Dina Tsagari and Ildiko Csépes (eds) (2012). *Collaboration in language testing and assessment*. Brussels: Peter Lang, 171-184.

- **Garant, Mikel** (2009). "A case for holistic translation assessment." *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 1, 5-17.

- **ISO 17100** (2015). *Translation services -- Requirements for translation services.*

- **Kockaert, Hendrik and Winibert Segers** (2012). "L'assurance qualité des traductions : items sélectionnés et évaluation assistée par ordinateur." *Meta: Translators' Journal* 57(1), 159-176.

- --- (2014). "Evaluation de la traduction : la méthode PIE (Preselected Items Evaluation)." *Turjuman* 23(2), 232-250.

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

- **Matlock-Hetzel, Susan** (1997). "Basic Concepts in Item and Test Analysis." Paper presented at *Annual meeting of the Southwest Educational Research Association* (Austin, TX, 23-25 January, 1997).

- **McAlester, Gerard** (2000). "The Evaluation of Translation into a Foreign Language." Christina Schäffner and Beverly Adab (eds) (2000). *Developing Translation Competence*. Amsterdam: John Benjamins, 229-241.

- **PACTE** (2005). "Investigating Translation Competence: Conceptual and Methodological Issues". *Meta: Translators' Journal* 50(2), 609-619.

- --- (2009). "Results of the Validation of the PACTE Translation Competence Model: Acceptability and Decision Making." *Across Languages and Cultures* 10(2), 207-230.

- **Sabri, Shafizan** (2013). "Item Analysis of Student Comprehensive Test for Research in Teaching Beginner String Ensemble Using Model Based Teaching among Music Students in Public Universities." *International Journal of Education and Research* 1(12), 91-104.

- **Waddington, Christopher** (2001). "Different Methods of Evaluating Student Translations: The Question of Validity." *Meta: Translators' Journal* 46(2), 311-325.

- **Wiersma, William and Stephen G. Jurs** (1990). *Educational Measurement and Testing.* Boston: Allyn and Bacon.

Biographies

Hendrik J. Kockaert lectures French Linguistics, Terminology, Legal Translation, and Translation Technology at KU Leuven, Faculty of Arts, in Antwerp. Since August 2015, he is Dean of the Faculty of Arts on Campus Sint Andries in Antwerp. He is a Research Associate at the University of The Free State, Republic of South-Africa. He is a certified LICS Auditor for granting ISO 17100 certification to translation services, and he is the Editor-in-Chief of The Journal of Internationalisation and Localisation [JIAL]. He is the Chairperson of ISO TC 37 SC 1 and a member of NBN, the Belgian Standardization Institute. He is an expert in the development of ISO terminology standards. He is a certified ECQA (European Certification and Qualification Association) Terminology Manager Trainer and Job Role Committee Provider. He was the coordinator of the following projects financed by the European Commission: QUALETRA (JUST/2011/JPEN/AG/2975), and LIT Search (JUST/2013/JPEN/AG/4556). He publishes in the areas of terminology, translation quality assurance, and translation technology.
E-mail: hendrik.kockaert@kuleuven.be

Winibert Segers is affiliated to KU Leuven. He lectures on translation of administrative, cultural and medical texts. He publishes in the following research areas: translation evaluation, translation didactics, translation ethics, translation philosophy, translation theory and translation of travelogues.
E-mail: winibert.segers@kuleuven.be

Notes

[1] The terms 'score' and 'mark' must be distinguished. Marking is adapting a score. For example, a candidate gives a correct answer to five of the ten items. His/her score is 5/10, but his/her mark can be 3/10. A score can be adapted by using a z-score.

[2] Quality in Legal Translation (2012-2014) — JUST/2011/JPEN/AG/2975 (http://www.eulita.eu/qualetra).

[3] Trans-European Voluntary Certification for Translators (2013-2015) – Lifelong Learning Programme (http://transcert.eu/).

[4] This is a four step scheme for competence measurement: Step 1 competence and sub-competences definitions → Step 2 competence and sub-competences descriptors (Descriptors are 'can do' statements.) → Step 3 behavioral indicators (Behavioral indicators are objectively observable and dichotomously scorable.) → Step 4 measurement instruments.

[5] D index = p value Top Group minus p value Bottom Group (e.g. -0.06 for item 1).