Alireza Akbari* and Winibert Segers

Translation Difficulty: How to Measure and What to Measure

DOI 10.1515/les-2017-0002

Abstract: The present research opens up the theoretical light on measuring translation difficulty through various perspectives. However, accurate evaluation of translation difficulty by means of the level of the text, translator's characteristics, and the quality of translation are significant for translation pedagogy and accreditation. To measure translation difficulty, one has to scrutinize it into four ways as (1) the identification of resources of translation difficulty, (2) the measurement of text readability, (3) the measurement of translation difficulty by means of translation evaluation products such as holistic, analytic, calibrated dichotomous items (CDI), and the preselected items evaluation (PIE) methods, and (4) the measurement of mental workload. This article will expand on the mentioned factors in detail in order to shed light upon translation difficulty on how and what to measure.

Keywords: Translation difficulty, text readability, translation evaluation products, mental workload

1 Introduction

Understanding translation difficulty, especially in translation workshops, courses, and assignments, is regarded as the significant issue in translation pedagogy and research. Mostly, translation assignments and workshops are carried out holistically (based on impression) to gain the proposition of a text's difficulty. However, translation courses, particularly those with practical facets, require being evaluated more effectively and correspondingly the results more objectively. In the scope of foreign language teaching (FLT), researchers have indicated that assigning the suitable level of a task's difficulty improves the

***Kontaktperson:** Alireza Akbari, KU Leuven, Naamsestraat 63, 3000 Leuven, Belgien,
E-Mail: alireza.akbari@kuleuven.be

Winibert Segers, KU Leuven, Naamsestraat 63, 3000 Leuven, Belgien,
E-Mail: winibert.segers@kuleuven.be

academic results. In the field of translation assessment, the tasks of translation should be done in such a way that they are designed in an order of difficulty for trainee translators. In this respect, the complexity and the difficulty of the source text [language] must be elucidatory and suitable for them. Thus, there is an obligation for texts to be leveled in the pedagogy of translation assessment.

Other complexities arise in the field of translation studies, making translation difficulty a worthy and necessary exploration. For example, translation researchers mostly have no excellences to allude to when they choose passages to be tested in terms of the type of the texts, length, and difficulty in process-oriented research (Krings 2001:74). This situation makes translation assessment and evaluation more challenging. For instance, the utilization of diverse translation strategies in terms of the type of the text might differ depending on the translation difficulty levels of the intended text. According to Dragsted (2004), the skillful renderers would designate a more novice-like behavior when assessing and evaluating translation of the difficult text. Therefore, paying attention to the role of translation difficulty in terms of the process of translation, text types, length, and the frequency of translation strategies applied by the renderer is of critical significance. Unfortunately, there has been little research on translation difficulty (Sun 2015); however, Jensen (2009) is among the researchers carrying out empirical research within this field. In line with Jensen, Campbell and Hale (2003) maintain that combining ‘an inventory of universal sources of translation difficulty’ in the source texts and ‘inventories for specific pairs of language’ is attainable.

The aim of this research is to scrutinise translation difficulty theoretically. However, before examining translation difficulty, one needs to ascertain how to measure difficulty and correspondingly what to measure.

2 The Notion of Translation Difficulty

Cognitively speaking, difficulty alludes to the mental endeavor of solving a problem and the amount of cognitive energy spent in the endeavor. With this in mind, translation difficulty can be viewed as the cognitive resources utilized by the tasks of translation particularly for the renderer (translator) to scrutinize the subjective and objective criteria. Dahl (2004:390) argues that translation difficulty is primarily related to the tasks and also pertains to person; therefore, ‘difficulty’ implies information objectivity and can be calculable. In order to evaluate and assess the tasks (translation tasks) and to make them appropriate based on the abilities, restrictions, and the expectation of people (Karwowski 2006), ‘mental workload’ has been an important subject in the scope of ergonomics (Moray

1979). Up to now, there has not been an exhaustive definition of mental workload. However, according to Gopher and Donchin (1986:41), this term refers to

Difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time.

According to the model proposed by Meshkati (1988), causal factors and effect factors play a significant role in mental workload. The former refers to the task and environmental variables involving the volume of information, task urgency, structure and the inadaptability of the tasks, the pressure of time, task originality, and task frequency. The latter alludes to operator's characteristics and moderating variables consisting of individual cognitive strengths, motivational and personal states, and various experiences of the past. On balance, the causal and the effect factors are reciprocally affiliating with each other.

Psychologically speaking, cognitive load is defined as 'the demand for working memory resources required for achieving goals of specific cognitive activities in certain situations' (Kalyuga 2009:35). However, scholars and researchers themselves express various notions about cognitive load. According to Sammer (2006), cognitive load can be considered as the sub-category of mental workload. As such, cognitive load theory is concerned with the aftermaths of learning prediction and the restrictions of human cognition (Plass et al. 2010).

In light of these explanations, task difficulty, mental workload, and cognitive load are similar and they consider the 'information processing demands that a certain task imposes on an individual and the limit of working memory' (Sun 2015:32). However, to juxtapose these three terms with one another, the theories of mental workload are far more related to the expression 'translation difficulty'. Therefore, the present research applies the term translation difficulty for the following reasons: (1) the term difficulty is more prevalent compared to the term 'workload' (Nord 2005 and Robinson 2011) and (2) translation difficulty underscores the hallmarks of the task aiming at measuring objectively.

3 Translation Difficulty: Resources

In light of the mentioned explanations, translation difficulty resources can be categorized into two groups as: (1) Factors related to translation (task) and (2) Factors related to translator (renderer). These two factors are mutually inclusive and need to be evaluated simultaneously.

3.1 Task Factors

Process-oriented translation and the tasks of translation involve decoding, analyzing, and apprehending on the one hand, rewriting, encoding, re-contextualizing, and recoding on the other hand (Wilss 1982). In doing so, the renderer first decodes (reads) the source text and secondly encodes it in the reciprocal language (target). The process of reading, and in particular the reading apprehension (Kamil et al. 2011), is an especially pertinent factor to translation difficulty. Task factors involve text readability and difficulty in the translation of specific items.

3.1.1 Text Readability

Text readability is regarded as the significant issue in translation difficulty. According to Reading Study Group (RAND) (2002:25), the following dimensions from reading confront readers:

- Discourse genre such as narration, description, exposition, and persuasion;
- Discourse structure including rhetorical composition and coherence;
- Media forms such as textbooks, multimedia, advertisements, hypertext, and the Internet;
- Sentence difficulty including vocabulary, syntax, and the propositional text base (the explicit meaning of the text's content drawn from propositions in the text, i.e. statements or idea units, but without more-subtle details about verb tense and deictic references);
- Content including different types of mental models, cultures, and socio-economic strata; age-appropriate selection of subject matter; and the practices that are prominent in the culture;
- Texts with varying degrees of engagement for particular classes of readers.

In this vein, text factor can be grouped as syntactic and lexical difficulty and content and subject matter. Syntactic and lexical complexity allude to '*the number of sundry intricate words*', '*the number of simple words*', '*the percentage of monosyllables*', '*the number of personal pronouns*', '*the average sentence length in words*', '*the percentage of diverse words*', '*the number of propositional phrase*', and lastly '*the percentage of simple sentences*' (Gray and Leary 1935:16). Therefore, the afore-mentioned factors 'lend themselves most readily to quantitative enumeration and statistical treatment' (ibid.). Subject matter and content both refer to the non-concrete texts and tellingly they are completely difficult to perceive compared to concrete texts elucidating real events, phenomena, and activities. In this

direction, some questions (Nord 2005:98) may help in understanding the critical information about the subject matter:

1. Is the source text a thematically coherent single text or a text combination?
2. What is the subject matter of the text (or of each component of the combination)? Is there a hierarchy of compatible subjects?
3. Does the subject matter elicited by internal analysis correspond to the expectation built up by external analysis?
4. Is the subject matter verbalized in the text (e.g. in a topic sentence at the beginning of the text) or in the text environment (title, heading, sub-title, introduction, etc.)?
5. Is the subject matter bound to a particular (SL, TL, or other) cultural context?
6. Do the TC conventions dictate that the subject matter of the text should be verbalized somewhere inside or outside the text?

The term ‘content’ or ‘meaning’ is constrained more or less to the lexical levels. Therefore, this term remains indeterminate and there are not many ideas on how to exploit the meaning (content) of the text (Nord 2005). Content and its role in the text mostly appear in the figure of ‘paraphrase’ (Bühler 1984). In this respect, however, whenever a translator masters the source language text, the rules, conventions, and the norms governed in the text, he/she no longer encounters any problems related to the content of the text.

3.1.2 Translation-specific Problems

In the scope translation assessment, Nord (2005) categorizes translation problems into four groups in that every renderer has to find some solutions regardless of his/her competence level: (1) problems of text specific translation, (2) problems of pragmatic translation such as text recipient, (3) problems of cultural translation such as styles and conventions of the text, and (4) problems of linguistic translation as verb tense usages across languages. In line with Nord, Gregory Shreve et al. (2004) posit the following factors as (1) textual variance, (2) textual degradation such as fragmentary texts, (3) linguistics distance between SL and TL, (4) cultural distance between SL and TL, (5) lexical intricacy (complexity), (6) syntactic complexity, and (7) propositional complexity. According to Ervin and Bower (1953), translation difficulty can originate from three factors putting a translator into a blurred situation. They are (1) ‘lexical meaning’ (such as objective referents, homonyms, affective and figurative meaning, untranslatable concepts, and similar words), (2) ‘grammatical meaning’ (e.g. syntactical requirements, and stylistic factors), and (3) the ‘functional equivalence in different cultural contexts’. How-

ever, what discriminates translation difficulty from text difficulty can be watered down to one item as ‘equivalence’ and its role in theoretical and practical translation. According to Pym (2010:7), equivalence ‘is a relation of equal value between the source text segment and a target text segment and can be established on any linguistic level from form to function’. For instance, Otto Kade (1968) categorizes equivalence into ‘one to one’ (Eins-zu-Eins) for technical terms and concepts, ‘one to several’ (Viele-zu-Eins; *Facultativ* in German) opting from the alternatives, ‘one to part’ (Eins-zu-Teil; *Approximativ* in German) for partial matches of the available equivalents, and ‘one to none’ (Eins-zu-Null) for neologisms. In this direction, the clear-cut difficulty status of translation tasks includes competency and translator competency explained in more detail in the next section.

3.2 Translator Factors

Factors which are usually related to the role of translator between the source and target languages are considered translation competence. The term competence in translation is defined as ‘a complex know how to act resulting from integration, mobilization, and organization of a combination of capabilities and skills (which can be cognitive, affective, psycho-motor or social) and knowledge (declarative knowledge) used efficiently in situations with common characteristics’ (Lasnier 2000). In addition, European Master’s in Translation (2009:3) defines competence as ‘the combination of aptitudes, knowledge, and behavior and know-how necessary to carry out a given task under given condition’. Until now, a number of competence categories have been submitted; however, the most updated ones are related to PACTE (2003), Albir (2007), and EMT (European Master’s in Translation) (2009). PACTE (2003) advances five sub-competences as (1) ‘bilingual sub-competence’ (e.g. pragmatic, textual, sociolinguistic, and lexico-grammatical knowledge), (2) ‘extralinguistic sub-competence’ (e.g. thematic and bi-cultural knowledge), (3) ‘knowledge about translation sub-competence’ (e.g. processes, methods, procedures, translation briefs, and users), (4) ‘instrumental sub-competence’ (documentation resources and the applied technologies), and (5) ‘strategic sub-competence’ (planning and evaluating the process of translation, identifying translation problems, and implementing the relevant procedures to solve the problems). The translation competence model proposed by Albir (2007) is divided into six categories: (1) ‘methodological and strategic competences’ (informative target of teaching translation, textual equivalence, and process of translation), (2) ‘contrastive competences’ (translation from language A to B and B to A and the role of target language in translation), (3) ‘extralinguistic competences’ (the importance of extralinguistic features in translation such as institutions, author,

target reader, customer, publisher, markets, etc., the importance of encyclopedia and the thematic knowledge to solve translation problems, and also the importance of translation techniques based on the genre and the receiver of the text such as glossing, footnotes, and parenthesis), (4) ‘occupational competences’ (e.g. the importance of translation job markets), (5) ‘instrumental competences’ (e.g. documentation resources (either paper-based or digital based) and parallel texts), and (6) ‘textual competences’ (e.g. typical text features and text genres). And consequently, in line with Albir (2007), EMT (2009) also proposes six categories of competences based on group experts as (1) ‘translation service provision competence’ (interpersonal dimension and production dimension), (2) ‘language competence’ (understanding grammatical, lexical, idiomatic structures, and typographic conventions and utilizing specific structures in both languages), (3) ‘intercultural competence’ (sociolinguistic dimension and textual dimension), (4) ‘information mining competence’ (using search engines and tools such as IATE, EU-JRC Acquis Multilingual Corpus, and DGT Translation Memory, utilizing strategies for terminological research, and enhancing some criteria for evaluating the reliability of documentary resources), (5) ‘thematic competence’ (understanding how to search for suitable information to gain a better insight of the thematic facets of a document), and finally (6) ‘technological competence’ (managing the databases and files, familiarizing with new tools, and producing a translation in various formats for different technical media). However, various tasks in translation generally require adopting the certain competence of translations since every translator has a genuine combination of the mentioned competences and sub-competences.

4 How to Measure Difficulty in Translation

This section deals with the measurement of translation difficulty in two steps as text difficulty (readability) and the elicitation of translation problems.

4.1 Text Difficulty

Text difficulty or readability is the facility of a reader to perceive a written text. Readability formula also is the equation combining the ‘statistically measurable text features’ (Sun 2015) which can predict the difficulty of text such as the number of the words, number of sentences, number of syllables, the number of characters, and so on. Of the proposed readability formulas, Flesch Reading Ease, Dale-Chall Readability Score, Flesch Kincaid Readability Score, Coleman-Liau

Readability Score, Bormuth Readability Score, SMOG Index, Spache Readability Index, and Fry Graph and Raygor Readability Tool are among the customary ones. To understand the function of these formulas, this research paper is going to explain them in detail. The first readability formula is Flesh Reading Ease applied for the sentence length and the number of syllables per word. This formula is applied for evaluating the grade level of the reader (Flesch 1948).

Flesch Reading Ease:

$$206.835 - \left(1.015 \times \frac{\text{number of words}}{\text{number of sentences}} \right) - \left(84.6 \times \frac{\text{number of syllables}}{\text{the number of words}} \right)$$

The end scores ranges from zero to one hundred. The scores between 90.0 and 100.0 are considered as the easiest text to read (an average 5th grader); while the scores from 0.0 to 30.0 as the highest difficulty of text, understood by the college graduate.

The second formula is tagged as Dale-Chall Readability Score utilized for word length to evaluate word difficulty (Dale-Chall 1948). It calculates the US grade level based on sentence length and the number of difficult words.

Dale-Chall Readability Score:

$$0.1579 \left(\frac{\text{difficult words}}{\text{words}} \times 100 \right) + 0.0496 \left(\frac{\text{words}}{\text{sentences}} \right)$$

This formula is considered accurate since it is based on the use of the familiar words in the text, rather than scrutinizing the syllable and the letter counts.

Table 1: Dale-Chall Readability Formula (1948)

Adjusted Score	Grade Level
4.9 and Below	Grade 4 and Below
5.0 to 5.9	Grades 5–6
6.0 to 6.9	Grades 7–8
7.0 to 7.9	Grades 9–10
8.0 to 8.9	Grades 11–12
9.0 to 9.9	Grades 13–15 (College)
10 and Above	Grade 16 and Above (College Graduate)

Another popular formula is Flesch Kincaid Formula proposed by Peter Kincaid et al. (1988). It is based on the average number of syllables per word and words per sentence for technical documents.

Flesch Kincaid Readability Score:

$$\left(0.39 \times \frac{\text{number of words}}{\text{number of sentences}}\right) + \left(11.8 \times \frac{\text{number of syllables}}{\text{number of words}}\right) - 15.59$$

The other formula is labeled as Coleman-Liau Readability Score, calibrating the readability of all text books for the public school system. This readability index is on the basis of the characters (length of the characters) instead of syllables per word. Also, this formula relies upon the calculation of the samples of the hard copies. Moreover, one can scan the hard copy of the text using a free Optical Character Recognition (OCR) program to identify the word, character, and sentence and exerts the intended formula to the text. Based on Coleman (1975), ‘there is no need to estimate syllables since word length in letters is a better predictor of readability than word length in syllables’.

Coleman-Liau Readability Score:

$$\left(5.89 \times \left(\frac{\text{number of characters}}{\text{number of words}} / \frac{\text{number of words}}{\text{number of sentences}}\right)\right) -$$

$$\left(\left((3 \times \text{average number of sentences}) / 1000\right) \times \frac{\text{number of words}}{\text{number of sentences}}\right) - 15.80$$

Bormuth Readability Score investigates the readability of text based on two factors: (1) the average length of the characters and (2) the average number of the familiar words in the original. However, the intended formula applies for the texts above 4th grade level (Bormuth 1966). The differences between Dale-Chall Readability Score and Bormuth formula depend on the character count instead of syllable count, and the average number of familiar words in the original instead of the percentage of difficult words.

Bormuth Readability Score:

$$\begin{aligned} &0.886593 - \left(\frac{\text{number of characters}}{\text{number of words}} \times 0.03640 \right) + \\ &\left(\frac{\text{number of words in the original}}{\text{number of words}} \times 0.161911 \right) - \left(\frac{\text{number of words}}{\text{number of sentences}} \times 0.21401 \right) \\ &- \left(\frac{\text{number of words}}{\text{number of sentences}} \times 0.000577 \right) - \left(\frac{\text{number of words}}{\text{number of sentences}} \times 0.000005 \right) \end{aligned}$$

The next readability formula would be SMOG (Simple Measure of Gobbledygook) Index (McLaughlin 2007) inspecting the number of polysyllables and sentences. The SMOG index calculates a 0.985 correlation along with a SE (standard error) of 1.5159 grades. SMOG formula would be usable in cases in which the participants have at least 12 years of schooling (ibid). Meanwhile, the mentioned index is normed on thirty-sentence samples.

SMOG Index:

$$1.0430 \sqrt{(\text{number of polysyllables} \times \frac{30}{\text{number of sentences}})} + 3.1291$$

Table 2: SMOG Index (McLaughlin 2007)

SMOG Conversion Table	
Total Polysyllabic Word Count	Approximate Grade Level
1–6	5
7–12	6
13–20	7
21–30	8
31–42	9
43–56	10
57–72	11
73–90	12
91–110	13

Table 2: (continued)

SMOG Conversion Table	
Total Polysyllabic Word Count	Approximate Grade Level
111–132	14
133–156	15
157–182	16
183–210	17
211–240	18

The next readability formula is tagged as SPACHE Readability Index which makes use of the sentence length and the number of unique unfamiliar words in the text (Spache 1953). This formula is best used when text difficulty falls at the third grade level or below. Furthermore, it is utilized for primary school level texts through to the end of third grade.

Spache Readability Index:

$$\left(0.141 \times \frac{\text{number of words}}{\text{number of sentences}}\right) + (0.086 \times \text{Percentage of Unique Unfamiliar Words}) + 0.839$$

And finally, the last readability tools are regarded as Fry Graph and Raygor Readability Tools (1989). They are based on materials from primary and secondary schools in graph format plotting the average number of sentences (Y axis) against the average number of syllables (X axis) per 100 words. In this direction, the intersection of Y axis to X axis shows the reading difficulty. For instance, the graphs below show the degree of difficulty of the selected text through Fry and Raygor Readability Tools.

Figure 1: Fry and Raygor Graphs

5 How to Measure Translation Difficulty Based on Translation Evaluation Products

5.1 Translation Evaluation

The target of Translation Evaluation is to scrutinize and mark students' drafts and translators' competences, analyze linguistic and paralinguistic facets of the text, peruse tools and models of translation evaluation, develop an independent method to retrieve data which quantifies scientific output expatiating societal impacts, and value and validate translation evaluation results of societal quality on the basis of specific translation theory and competence. Juxtaposition of the source text with target text and also target text per se are the processes in translation evaluation (Mobaraki and Aminzadeh 2012:63). In translation evaluation, translation product as translated text and the end-product of the translation process, translation process, mental states and sub-tasks, situatedness (Muñoz Martín, 2010), the translation services (e.g. agents, players, etc.), real world resources (e.g. parallel texts and corpora and different modes of testing), and finally translator's competence (e.g. linguistic, cultural, research and technical competences) (Kockaert and Segers 2016) are applied.

Via translation evaluation, one can determine, on one hand, the suitability, readability, and usability of the translated text, and on the other hand, cultural, linguistic, and technical competences of the translator. In this direction, McAles-ter (2000:230), proposes four criteria in translation evaluation as readability, validity, subjectivity, and practicality.

5.2 Translation Grading

Translation grading is a means to understand the translation problem. In this direction, there exist methods scrutinizing and measuring translation in professional settings. Holistic, analytic, CDI (calibration dichotomous items), and PIE (preselected items evaluation) are methods categorized under translation grading.

5.2.1 Holistic Method

Recently, assessment has become the prevalent issue within the domain of translation studies. A number of research studies on translation quality assessment (Bowker 2001; Melis and Albir 2001; Williams 2001) provide the situation about different methods and models of translation evaluation. Of these methods, the holistic method is of great importance in subjective evaluation. Holistic method is a mode of evaluating the quality of translation by giving a score on the basis of the translator's overall impression (Mariana, Cox, and Melby 2015). Holistic or the 'intuitive-impressionistic' method (Eyckmans, Anckaert, and Segers 2009) is the prevalent utilized method in professional settings. Generally, in order to evaluate a translation holistically, the evaluator uses his/her overall intuition for scoring the intended translation as excellent, good, fair, or very bad. In this respect, the evaluator securitizes the end product as a whole without considering the errors. Therefore, the holistic method is fully subjective and mostly related to the taste of the evaluator. Tellingly, the holistic method is found to vary from one teacher/evaluator to another teacher/evaluator. According to Kockaert and Segers (2014), 'a holistic approach seemed to focus better on a context sensitive evaluation, and seemed to discard exclusive attention to grammatical errors in translation tests.' Context sensitive evaluation signifies that 'translations do not occur in a vacuum, that they need to be interpreted and evaluated in their relevant context' (Koskinen 2008:72). Noticingly, McAlester (2000:231) came to a similar conclusion that 'often the actual evaluation follows fairly rough guidelines based admittedly in the best cases on experience and common sense, but in the worst on mainly subjective impressions'. However, the reader should pay attention not to label the holistic method as an unsystematic one. Garant (2009:10) argues that:

Holistic method refers to a systematic way in which the teacher arrives at an overall impression of the text as opposed to relying on a discrete point-based scale. The teachers in that group had each devised their own, systematic way of evaluating translations.

5.2.2 Analytic Method

Analytic method is a method to evaluate the quality of translation by perusing the segment of the text (e.g. sentence, paragraph, individual words, etc.) based on certain criteria rather than focusing on the evaluator's impression. Analytic method not merely inspects the error counting, 'but also assesses and characterizes to the extent that the two most common criteria for this characterization are nature and importance' (Conde 2011:70). It is claimed that the analytic method based on error analysis is 'more reliable and valid than holistic method' (Waddington 2001:36).

In translation studies, error distinction according to its importance has received much research attention (Ceschin 2004; Darwish 2001; Rosenmund 2001; Vollmar 2001; and Koo and Kind 2000). Hierarchically, there stand major errors at the top and minor errors at the bottom. In this vein, error's effect or impact on the whole text is of remarkable significance (Cruces 2001:816; Martinez and Hurtado 2001; and Larose 1998:16).

According to Eyckmans, Anckaert, and Segers (2013), the translation errors must be marked based on evaluation grids provided below. In this direction, the evaluator must underline every error (both considering language errors and translational errors) and consequently maintain the relevant information in the margin owing to the nature of the errors.

Table 3: Analytic Method (Adopted from Eyckmans, Anckaert, and Segers 2013)

SENS (Meaning or Sense)	Toute altération du sens dénotatif: informations erronées, non sens..... J'inclus dans cette rubrique les oublis importants, c'est-à-dire font l'impasse sur une information d'ordre sémantique. <i>(Any deterioration of the denotative sense: erroneous information, nonsense, important omission.....)</i>	-1
CONTRESENS (Misinterpretation)	L'étudiant affirme le contraire de ce que dit le texte: information présentée de manière positive alors qu'elle est négative dans le texte, confusion entre l'auteur d'une action et celui qui la subit..... <i>(The student misinterprets what the source text says: information is presented in a positive light whereas it is negative in the source text, confusion between the person who acts and the one who undergoes the action.....)</i>	-2
VOCABULAIRE (Vocabulary)	Choix lexical inadapté, collocation inusitée <i>(Unsuited lexical choice, use of non-idiomatic collocations)</i>	-1
CALQUE (Calque)	Utilisation d'une structure littéralement copiée et inusitée en français <i>(Cases of literal translation of structures, rendering the text into French)</i>	-1

Table 3: (continued)

REGISTRE (Register)	Selon la nature de texte ou la nature d'un extrait (par exemple, un dialogue): traduction trop (in)formelle, trop recherchée, trop simpliste.... <i>(Translation that is too (in)formal or simplistic and not corresponding to the nature of the text or extract)</i>	-0.5
STYLE (Style)	Lourdeurs, répétition maladroits, assonances malheureuses..... <i>(Awkward tone, repetition, unsuited assonances)</i>	-0.5
GRAMMAIRE (Grammar)	Erreurs grammaticales en français (par exemple, mauvais accord du participe passé, confusion masculine/féminin, accords fautifs....) + mauvaise compréhension de la grammaire du texte original (par exemple, un passé rendu par un présent...) et pour autant que ces erreurs ne modifient pas en profondeur le sens. <i>(Grammatical errors in French (for example, wrong agreement of the past participle, gender confusion, wrong agreement of adjective and noun, ...) + faulty comprehension of the grammar of the original text (for example, a past event rendered by a present tense,), provided that these errors do not modify the in-depth meaning of the text)</i>	-0.5
OUBLIS (Omission)	Voir SENS <i>(See sense/ meaning)</i>	-1
AJOUTS (Addition)	Ajout d'informations non contenues dans le texte (sont exclus de point les étirements stylistique. <i>(Addition of information that is absent from the source text (stylistic additions are excluded from this category))</i>	-1
ORTHOGRAPHE (Spelling)	Erreurs orthographiques, pour autant qu'elles ne modifient pas le sens. <i>(Spelling errors, provided they do not modify the meaning of the text)</i>	-0.5
PONCTUATION (Punctuation)	Oubli ou utilisation fautive de la ponctuation. Attention: l'oubli, par exemple, d'une virgule induisant une compréhension différente du texte, est considéré comme une erreur de sens. <i>(Omission or faulty use of punctuation. Caution: the omission of a comma leading to an interpretation that is different from the source text, is regarded as an error of meaning or sense)</i>	-0.5

Another point regarding error analysis is toward the location of errors in the analytic method. Vollmar (2001:26) regards errors as leading to the misinterpretation of 'significant portion' of the source text. Moreover, Hadjú (2002:249) considers location of error to the parts of the text of which they are particularly noticeable, such as in section headings. House (2001:151) considers location of errors in 'titles, addresses, phone numbers, and indexes' as major errors.

In comparison to the holistic method, analytic method requires devoting more time to the detail of the text. However, in analytic method, the translator has better insight into the correctness of translation.

In spite of this, the analytic method has one demerit as ‘the look at the target text as a whole is sometimes lost’ (Kockaert and Segers 2016). The analytic method, arguably, is also subjective since different evaluators or graders are not always in the position of agreement: perhaps, one slight error for one evaluator is regarded as a major error for another evaluator. Henceforth, to solve the subjectivity of the evaluation, CDI and PIE methods have been proposed.

5.2.3 Calibrated Dichotomous Items (CDI) Method (Norm-referenced Method)

Translation studies need to be implemented empirically for the quality of translation tests (Waddington 2004). However, it seems that assessment in translation is based on the code of practice rather than scrutinizing empirically and therefore, one can label it as criterion referenced assessment. Criterion-referenced assessment alludes to items that are ‘based on specific objectives, or competency statements’ (Shrock and Coscarelli 2007:25). The items are not planned to distinguish one score from another, but to determine whether a person is able to perform a certain objective (Mariana, Cox, and Melby 2015).

Researchers and scholars in the field of translation studies (PACTE 2000; Conde 2005) are now preparing the situation for issues such as inter- and intra-reliability of translation assessment (Eyckmans, Anckaert, and Segers 2013:513). In this vein, norm-referenced methods for evaluating translation with the purpose of freeing translation both analytically and holistically have come onto the horizon. According to Schrock and Coscarelli (2007:25), norm-referenced assessment is ‘composed of items that will separate the scores of the test-takers from one another’. Therefore, norm-referenced assessment in translation is responsible for ranking a group of translators from best to worse according to some pre-established criteria.

The Calibration Dichotomous Items (CDI) method is one such assessment, based on the practice of calibrating segments of translation and norm-referenced assessments (Eyckmans, Anckaert, and Segers 2009:76). The calibration of segments in translation allows the ‘construction of the standardized test’ (ibid).

CDI method is tagged as a norm referenced method since it is ‘independent of subjective a priori judgments upon the source text and the target translated text’ (Eyckmans, Anckaert, and Segers 2013). Every component of the translated text leading to make differences is labeled as an ‘item’ and it is selected in a pre-test procedure. As its name clears, CDI assumes the text segments dichotomously, i.e.

a segment is either correct or not. The dichotomy of segments does not allude to inappropriateness of translation; however, it signifies that for each translated segment the graders are in the position of agreement whether or not one alternative is acceptable/correct.

In order to identify text segments, discriminating power is the right option in CDI method. According to Ferrando (2010:111), discriminating power can be organized on the basis of three criteria: '(1) the type of score, (2) the range of discrimination, and (3) the conceptualizations and aspects that are measured'. According to Lord (1980) and McDonald (1999), discriminating power can be measured for all types of scores. However, in this respect, some evaluators concentrate solely on single-item scores or 'total-test scores' provided by sum of the item score (Levine and Lord 1995). It means that the researcher compares the number of participants (pilot study) along with high test scorers who correctly answered the item with those of low scorers who correctly answered the same item as upper-lower index (Lord 1980).

CDI method does not institutionalize a time-saving process; hence, this method will be full-fledged utilizable in summative tests. This is because a summative assessment is a judgment (decision) covering all the 'evidence up to a given point' (Taras 2005:468). Therefore, objectivity is the mere trait of the CDI method since it is an 'evaluator independent measurement' (Eyckmans, Anckaert, and Segers 2009:76) filling the gap between theories of testing and characteristics of translation studies.

5.2.4 Preselected Items Evaluation (PIE) Method

The preselected items evaluation (PIE) method is the feasible version of CDI method in translation evaluation (Anckaert, Eyckmans, Justens, and Segers 2013; Kockaert and Segers 2012, 2014). PIE method is a system designed by Hendrik Kockaert and Winibert Segers which is highly suitable for summative assessment. PIE method is based on the number of preselected items constrained in evaluation.

To understand the function of PIE method in translation assessment, one has to be concerned about the role of multidimensional quality metrics (MQM) as a new framework for translation quality assessment (Mariana, Cox, and Melby 2015). This is because the application of MQM is based on the preselected items evaluation (PIE) method. The remarkable traits of the intended paradigm are as follows: (1) it is an exhaustive paradigm for translation quality assessment; (2) all metrics in this paradigm draw on the same hierarchy of errors; (3) it is customizable to the needs of the user; and (4) each metric in this paradigm is related to

some sets of specifications which are fundamental for elaborating on the standards of quality (ibid).

As another point in translation evaluation, PIE method is also a calibrated and dichotomous method (both norm-referenced and criterion-referenced). Henceforth, the preselected items in PIE method are chosen on the basis of item difficulty (p-docimology) and item discrimination (d-index) for which the 'correct and erroneous solutions are determined' (Kockaert and Segers 2016).

P-docimology: Item Difficulty

According to Rasinger (2008:159), 'significance in statistics refers to the probability of our results being a fluke or not; it shows the likelihood that our result is reliable and has not just occurred through the bizarre constellation of individual numbers.' In this direction, statistical significance can be argued in terms of 'p docimology' tracing the 'probability that the results happened by chance' (Saldanha and O'Brien 2013:199). Therefore, the lower the p-docimology in translation evaluation, the higher confidence one has that the results did not occur haphazardly. Both CDI and PIE involve the calculation of item difficulty and item discrimination for every item in CDI and for the number of preselected items in PIE method. To put it in another way, p-docimology alludes to the percentage of participants who answer the item correctly.

D-index: Item Discrimination

According to Ary, Jacobs, Sorenson, and Razavieh (2010:211) 'the item discrimination index shows the extent to which each item discriminates among the respondents in the same way as the total score discriminates.' This can be done via correlating item scores with total scale scores. If the high scorers on an individual item have the high total scores and if the low scorers have the low total scores, therefore, the item is discriminating in the same way as the total score. In order that the item discrimination is to be useful, the item must correlate at least 0.25 with the total score. In this direction, items having low correlation or negative correlation with the total score should be deleted due to the fact that they are not contributing to the measurement of the attitude (ibid). Positive item discrimination is productive unless 'it is so high that the item merely repeats the information provided by other items on the test' which is called the 'attenuation paradox' (Ebel 1979).

In order to calculate the discrimination power of an item, two indices will be contributing: (1) discrimination index (D) and (2) discrimination coefficient. Ebel (1979) argues that:

The discrimination index (D) is computed from equal sized high and low scoring groups on the test. Subtract the number of successes by the low group on the item from the number of successes by the high group, and then divide this difference by the size of the group. The range of this index is +1 to -1. Using Truman Kelley's '27 % of sample' group size, values of 0.40 and above are regarded as higher and lesser than 0.20.

On the other side of the coin, discrimination coefficient involves every single person taking the test in spite of the fact that only 'the upper (27 %) and the lower scorer (27 %)' are included in the discrimination index (Sabri 2013:4). Wiersma and Jurs (1990:145) argue that '27 % is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis.'

To summarise, the PIE method is an attempt at figuring out goals such as the ideal length of the source text (number of the words), the ideal number of preselected items in the source language, the ideal p-docimology (item difficulty) (0.20–0.90 or 0.27–0.79?), the ideal norm values for the d-index (item discrimination), the utlizability of Ebel norms, the criteria for the preselection of the items, the reliability and validity of PIE method (construct validity, content validity, predictive validity, ecological validity, etc.) for the evaluation of translation products, and the automated evaluation of translation products on the basis of the evaluation part (EvaluationQ) of the software tool TranslationQ (Televic-KU Leuven). According to Kockaert et al. (2016), TranslationQ can be divided into four major steps as 'TrainingQ', TestQ, EvaluationQ, and RevisionQ' in which each of step covers 'training management and immediate feedback', 'test management, PIE, and analytical evaluation', and correspondingly, 'translation flow and revision flow' respectively.

Last but not least, 'translation brief relevancy', 'domain specific', and 'test specific' criteria are three factors characterized in PIE method (Kockaert and Segers 2016). In this vein, 'right and wrong solutions are listed for each preselected item (e.g. grammar (structural points), spelling, vocabulary, style, and other aspects) in the source text of the test' (ibid.).

Table 4: Comparison of All Methods (Adopted from Kockaert and Segers 2016)

	Holistic	Analytic	CDI	PIE
Number of Items	Exhaustive	Exhaustive	Docimologically Relevant Items	Translation Brief
Evaluation	Global	Grids/Criteria	Grids/Criteria	Grids/Criteria
Dichotomous	–	–	✓	✓
Calibration	–	–	✓	✓

Table 4: (continued)

	Holistic	Analytic	CDI	PIE
Acceptance of Alternatives	Expected/ Unexpected	Expected/ Unexpected	Expected/ Unexpected	Expected/ Unexpected
EN 15038 Compatible	✓	✓	✓	✓
Interrater Reliability	–	–	✓	✓
Criterion Referenced	–	–	–	✓
Norm Referenced	–	–	✓	–

6 How to Measure Mental Workload

Mental workload and the way to measure mental workload have become a sensitive and important issue in the field. According to Gopher and Donchin (1986), ‘mental workload is a hypothetical construct that describes the extent to which the cognitive resources required to perform a task have been actively engaged by the operator’. Defining this term would not be the absolute license; however, there must also be some ways to measure it. Of many suggested techniques for measuring mental workload, there exist three categorizations for measuring workload as (1) physiological, (2) subjective, and (3) performance-based measures.

6.1 Physiological Measures

Physiological measures utilize the changes in the body to objectively measure the level of workload pertained to a task. Most research done in this area can be grouped into five areas measuring: (1) cardiac activity, (2) respiratory activity, (3) eye activity, (4) speech activity, and (5) brain activity (Miller 2001). Cardiac activity can be measured through blood pressure, heart rate, and heart rate variability. Respiratory activity can be measured via respiratory rate and oxygen concentration. Eye activity is measured through interval of closure, eye blink, and horizontal eye movement (HEM) (eye movement scanning for instrumental panel) (Hanking and Wilson 1998).

According to Brenner et al. (1994), speech can be measured through rate, pitch, jitter, loudness, and shimmer. However, three of the mentioned factors

(pitch, loudness, and rate) are the most affected ones in speech measurement. In order to measure brain activity, electroencephalogram (EEG)-‘a recording of electrical activity made from the scalp’ (De Waard 1996), and electrooculogram (EOG)-‘saccadic eye movement’ (Galley 1993) are used. The two other brain measurements are electromyogram (EMG) and electrocardiogram (ECG). The former alludes to the ‘task irrelevant facial muscles that are not required in the motor performance of a task’ (De Waard 1996) whilst the latter refers to cardiac measures, particularly heart rate variability (*ibid.*). So far, however, these two measures are not studied as extensively as possible. According to Blascovich (2004:881), physical measures are ‘online, covert, and continuous and can be linked in the time to the expected occurrence of the physiological state or process being indexed, thereby providing simultaneous evidence of the strength or operation of the state or process’.

6.2 Subjective Measures

According to Sun (2015), ‘subjective measures typically involve having participants judge and report their own experience of the workload imposed by performing a specific task’. In this direction, subjective measures are resilient for different people with different capabilities since it is related to ability, state, and attitude. Subjective measures are used frequently because of their ‘high face validity’ (Vidulich 1988). Subjective measures can be grouped as (1) unidimensional and (2) multidimensional ratings. The former is considered as the simplest since it has no intricate analysis techniques. The latter is regarded as time consuming and has a complex rating with three to six dimensions (De Waard 1996).

Unidimensional scale rating can be divided into Modified Cooper Harper Scale (MCH) and Overall Workload Scale (OW). MCH is ‘a 10-point unidimensional rating scale that results in a global rating of workload’ (Hill et al. 1992). MCH is utilized to measure cognitive and perceptual workload (Wilson and Eggemeier 2006). OW is defined as ‘a single, 20-step bipolar scale that is used to obtain this global rating. A score from 0 to 100 (assigned to the nearest 5) is obtained’ (Hill et al. 1992).

Conversely, the multidimensional rating scale can be categorized as (1) NASA Task Load Index (NASA-TLX) and (2) Subjective Workload Assessment Technique (SWAT). According to Hart and Staveland (1988), NASA-TLX utilizes six category dimensions to evaluate mental workload: (1) mental demand (deciding, thinking, recalling, searching, etc.), (2) physical demand (pulling, pushing, etc.), (3) temporal demand (time pressure owing to task elements), (4) effort (how the performance is accomplished), (5) performance (how the targets perform successfully),

and (6) frustration (insecure, stressed, discouraged, etc.). In this respect, each mentioned dimension is anchored by bipolar representation as poor/good and low/high. Based on Hart's elucidations (2006:907), NASA-TLX is 'being used as a benchmark against which the efficacy of other measures, theories, or models [of workload measurement] are judged'. SWAT marks to produce 'a single rating scale with interval properties' (Hill et al. 1992). Some studies substantiate that SWAT technique is utilizable in predicting the changes in workload (De Waard 1996; Wierwille et al. 1993). The theory behind this technique is that 'it gains insight into the mechanism of human information processing resources, together with the notion that it is possible to derive a model, by some rational procedure, that has greater validity than that of an arbitrarily chosen model' (Hendy et al. 1993).

6.3 Performance Measures

Performance in the field of mental workload refers to the effectiveness of fulfilling a specific task (Wilson and Eggemeier 2006). Primary and secondary measures are the assets to measure performance in mental workload. However, according to Yeh and Wickens (1988), the assumption behind the primary and secondary measures is that people have constrained resources. In this respect, primary task performance refers to the capability of mental workload to perform the main task (Rehmann 1955). Primary task performance can be utilized in the driving situations involving Time-to-Line crossing (TLC) and steering wheel movement (De Waard 1996). Consequently, based on Mulder's viewpoint (1979), 'the basic idea of a secondary task is that it measures the difference between the mental capacity consumed by a main task, and the total available capacity'. In this direction, the two particular mental workload indicators in performance measures in either primary task performance or secondary task performance are the time on task (speed) and the number of errors (accuracy) (O'Donnell and Eggemeier 1986).

7 Conclusion

The present article has scrutinized the sources of translation difficulty and the ways to measure them operationally. However, there have so far been few theoretical studies in the field of translation difficulty. In this direction, the present research has shed light on measuring translation difficulty by classifying them into four groups as (1) the sources of translation difficulty, (2) the measurements of translation (text) readability, and (3) the measurement of translation difficulty

by means of translation evaluation products such as holistic, analytic, CDI, and PIE methods, and lastly (4) the measurement of mental workload. Importantly, PIE method by utilizing Televic-KU Leuven software (TranslationQ) paves the way for assessing translation difficulty systematically since it scrutinizes the items difficulty and index discrimination which are the most important factors in determining the difficulty of translation. Research in the field of translation difficulty can enhance the concentration of translation processes and can help considerably to our perception of translation as process-oriented in terms of the connection among the behavior of the translator, the characteristics of a text, and the quality of translation. This research further contributes to and has implications for translation pedagogy and teaching, where translation difficulty should be carefully considered.

Reference

- Anckaert, Ph., Eyckmans, J., Justens, D. and 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. Le Dizez, W. Segers (Eds.). *Le bon sens en traduction* (pp. 79–93). Rennes: Presses Universitaires de Rennes.
- Ary, D., Jackobs, L. C., Sorenson, C. and Razavieh, A. (2010): *Introduction to Research in Education* (8th edition). California: Wadsworth.
- Biber, D. (1989): A Typology of English Texts. *Linguistics*, 27 (1), 3–44.
- Blascovich, J. (2004): Psychophysiological Measures. In: Lewis-Beck, M.S. Bryman, A. and Liao, T.F. (eds) *The Sage Encyclopedia of Social Science Research Methods*. Thousand Oaks, Calif.: Sage. Vol. 1. 881–883.
- Bormuth, J. R. (1966): Readability: A New Approach. *Reading Research Quarterly*, 1, 79–132.
- Bowker, L. (2001): Toward a Methodology for a Corpus Based Approach to Translation Evaluation. *Meta*, 46, 2.
- Brenner, M., Doherty, E. T., and Shipp, T. (1994): Speech Measures Indicating Workload Demand. *Aviation Space and Environmental Medicine*, 65(1), 21–26.
- Bühler, K. (1984): Karl Bühler's Theory of Language. *Proceeding of the Conference Held at Kirchberg*, Aug 26, John Benjamins.
- Campbell, S. and Hale, S. (2003): Translation and Interpreting Assessment in the Context of Educational Measurement. In: Anderman, G. M. and Rogers, M. (eds) *Translation Today: Trends and Perspectives*. Clevedon; Buffalo, N.Y.: Multilingual Matters, pp. 205–224.
- Ceschin, A. (2004): Memória de Tradução: Auxílio ou Empecilho? PhD Thesis. Pontifícia Universidade Católica do rio de janeiro.
- Coleman, M. and Liao, T. L. (1975): A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60, 283–284.
- Conde, T. (2011): Translation Evaluation on the Surface of Text: A Preliminary Analysis. *The Journal of Specialized Translation*, 15.
- Cruces, S. (2001): El Origen de los Errores en Traducción. Domingo Pujante Gonzalez et al. (eds.) (2001) *Écrire, Traduire et Représenter la Fête*. Valencia: Universitat de Valencia, 813–822.

- Dahl, Ö. (2004): *The Growth and Maintenance of Linguistic Complexity*. Amsterdam; Philadelphia: John Benjamins.
- Dale, E and Chall, J. (1948): A Formula for Predicting Readability. *Educational Research Bulletin*, 27, 11–28.
- Darwish, A. (2001): *Transmetrics: A Formative Approach to Translator Competence Assessment and Translation Quality Evaluation for the New Millenium*.
- De Waard, D. (1996): *The Measurement of Drivers' Mental Workload*. University of Groningen: Groningen.
- Dragsted, B. (2004): *Segmentation in Translation and Translation Memory Systems: An Empirical Investigation of Cognitive Segmentation and Effects of Integrating a TM-System into the Translation Process*. Unpublished PhD dissertation, Copenhagen Business School.
- Ebel, R. (1979): *Essentials of Educational Measurement*. Englewood Cliffs, NJ : Prentice Hall.
- EMT Exert Group (2009): *Competences for Professional Translators, Experts in Multilingual and Multimedia Communication*, Retrieved October 1. 2013, from: http://ec.europa.eu/dgs/translation/programmes/emt/key_documents/emt_competences_translators_en.pdf.
- Ervin, S. and Bower, R. (1953): Translation Problems in International Surveys. *Public Opinion Quarterly*, 16 (4), 1952–53.
- Eyckmans, J., Anckaert, Ph. and Segers, W. (2009): The Perks of Norm-referenced Translation Evaluation. In C. Angelelli, H. Jacobson (Eds.). *Testing and assessment in translation and interpreting studies* (pp. 73–93). Amsterdam: John Benjamins.
- Eyckmans, J., Anckaert, Ph. and Segers, W. (2013): Assessing Translation Competence. *Actualizaciones en Comunicación Social Centro de Lingüística Aplicada*.
- Ferrando, J. P. (2010): Assessing the Discriminating Power of Item and Test Scores in the Linear Factor Analysis Model. *Psicologica*, 33, 111–134.
- Flesch, R. (1948): A New Readability Yardstick. *Journal of Applied Psychology*, 32 (3), 221–233.
- Fry, E. B. (1989): Reading Formulas: Maligned but Valid. *Journal of Reading*, 32 (4). 292–297.
- Galley, N. (1993): The Evaluation of the Electrooculogram as a Psychophysiological Measuring Instrument in the Driver Study of Driver Behavior. *Ergonomics*, 36 (9), 1063–1070.
- Garant, M. (2009): A Case for Holistic Translation Assessment. *A Finlande Soveltavan Kielitieteen Tutkimuksia*, 1, 5–17.
- Gopher, D. and Donchin, E. (1986): Workload: An Examination of the Concept. In: Boff, K. R., Kaufman, L. and Thomas, J. P. (eds) *Handbook of Perception and Human Performance, Vol. II: Cognitive Processes and Performance*. New York: Wiley.
- Gray, W. S. and Leary, B. E. (1935): *What Makes a Book Readable*. Chicago, Ill.: The University of Chicago Press.
- Hajdú, P. (2002): “The New Hungarian Translation of Aristotle’s Poetics: When Translation and Commentary Disagree.” *Across Languages and Cultures*, 3 (2), 239–250.
- Hankins, T. C., and Wilson, G. F. (1998): A Comparison of Heart Rate, Eye Activity, EEG and Subjective Measures of Pilot Mental Workload during Flight. *Aviation Space and Environmental Medicine*, 69 (4), 360–367.
- Hart, S. G. (2006): NASA-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*. Santa Monica: Human Factors and Ergonomics Society. 904–908.
- Hart, S. G. and Staveland, L. E. (1988): Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock, P. A. and Meshkati, N. (eds) *Human Mental Workload*. Amsterdam: New York: North-Holland. 139–183.
- Hatim, B. and Mason, I. (1997): *The Translator as Communicator*. London: Routledge.

- Hendy, K. C., Hamilton, K. M., and Landry, L. N. (1993): Measuring Subjective Workload – When Is One Scale Better Than Many. *Human Factors*, 35 (4), 579–601.
- Hill, S. G., Lavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L., and Christ, R. E. (1992): Comparison of 4 Subjective Workload Rating-Scales. *Human Factors*, 34 (4), 429–439.
- House, J. (2001): How do We Know when a Translation is Good? Erich Steiner and Colin Yallop (Eds) (2001). *Exploring Translation and Multilingual Text Production: Beyond Content*. Berlin/New York: Mouton de Gruyter, pp. 127–160.
- Hurtado, A. (2007): Competence-based Curriculum Design for Training Translators, *The Interpreter and Translator Trainer*, 1, 163–195.
- Jensen, K. T. (2009): Indicators of Text Complexity. *Copenhagen Studies in Language*, 37, 61–80.
- Kade, O. (1968): *Zufall und Gesetzmässigkeit in der Übersetzung*, Leipzig.
- Kalyuga, S. (2009): *Managing Cognitive Load in Adaptive Multimedia Learning*. Hershey, PA: Information Science Reference.
- Kamil, M. L., Pearson, P. D., Moje, E. B. and Afflerbach, P. P. (eds) (2011): *Handbook of Reading Research, Volume IV*. New York: Routledge.
- Karwowski, W. (ed.) (2006): *International Encyclopedia of Ergonomics and Human Factors* (2nd ed.). Boca Raton, FL: CRC/Taylor and Francis.
- Kincaid, JP, Braby, R, and Mears, J. (1988): Electronic Authoring and Delivery of Technical Information. *Journal of Instructional Development*, 11, 8–13. doi:10.1007/bf02904998.
- Kockaert, H.J. and Segers, W. (2012): L'assurance Qualité des Traductions : Items Sélectionnés et Évaluation Assistée par Ordinateur. *Meta : journal des traducteurs / Meta: Translators' Journal*, 57 (1), 159–176.
- Kockaert, H.J. and Segers, W. (2014): Evaluation de la Traduction : la Méthode PIE (Preselected Items Evaluation). *Turjuman*, 23 (2), 232–250.
- Kockaert, H.J. and Segers, W. (2016): Evaluation of Legal Translation: PIE Method (Preselected Items Evaluation). *Journal of Specialized Translation*, (forthcoming).
- Kockaert, H. J., Segers, W., Wylin, B., and Verbeke, D. (2016): *TranslationQ: Automated Translation and Evaluation Process with Real-time Feedback*. KU Leuven, Televic Education.
- Koo, S. L and Kinds, H. (2000): A Quality-Assurance Model for Language Projects. Robert C. Sprung (Ed.) (2000). *Translation into Success. Cutting-edge strategies for going multi-lingual in a global age*. Amsterdam: John Benjamins, pp. 147–157.
- Koskinen, K. (2008): *Translating Institution: An Ethnographic Study of EU Translation*, Manchester: St Jerome Publication.
- Krings, H. P. (2001): *Repairing Texts: Empirical Investigations of Machine Translation Postediting Processes* (Koby, G., Shreve, G., Mischerikow, K. and Litzer, S. Trans.). Kent, Ohio: Kent State University Press.
- Larose, R. (1998): Méthodologie de L'évaluation des Traductions. *Meta*, 43 (2), 163–186.
- Lasnier, F. (2000): *Réussir la Formation par Compétences*, Montreal: Guérin.
- Levine, R. and Lord, F.M. (1959): An Index of the Discriminating Power of a Test at Different Parts of the Score Range. *Educational and Psychological Measurement*, 19, 497–503.
- Lord, F.M. (1980): *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale: LEA.
- Mariana, V., Cox, T., and Melby, A. (2015): The Multidimensional Quality Metrics (MQM) Framework: A New Framework for Translation Quality Assessment. *The Journal of Specialized Translation*, 23.
- Martínez, M. N. and Hurtado, A. (2001): *Assessment in Translation Studies: Research Needs*, *Meta*, 46 (2).

- McAlester, G. (2000): The Evaluation of Translation into a Foreign Language. In C. Schaeffner and B. Adab (eds.), *Developing Translation Competence* (pp. 229–242). *Benjamins Translation Library*, 38. Amsterdam: Benjamins.
- McDonald, R. P. (1999): *Test theory: A unified treatment*. Mahwah (NJ): LEA.
- McLaughlin, G. H. (2007): SMOG Grading: A New Readability Formula. *Journal of Reading*. [http://www.harrymclaughlin.com/SMOG_Readability_Formula_G._Harry_McLaughlin_\(1996\).pdf](http://www.harrymclaughlin.com/SMOG_Readability_Formula_G._Harry_McLaughlin_(1996).pdf)
- Melis, N. M. and Albir, A. H. (2001): Assessment in Translation Studies: Research Needs. *Meta*, 46 (2), 272–287.
- Meshkati, N. (1988): Toward Development of a Cohesive Model of Workload. In: Hancock, P. A. and Meshkati, N. (eds) *Human Mental Workload*. Amsterdam; New York: North-Holland, pp. 305–314.
- Miller, S. (2001): *Workload Measures*. Iowa: The University of Iowa Press.
- Mobaraki, M. and Aminzadeh, S. (2012): A Study on Different Translation Evaluation Strategies to Introduce an Eclectic Method. *International Journal of English Linguistics*, 2 (6).
- Moray, N. (1979): *Mental Workload: Its Theory and Measurement*. New York: Plenum Press.
- Moessner, L. (2001): Genre, Text Type, Style, Register: A Terminological Maze? *European Journal of English Studies* Vol. 5. No. 2. 131–138.
- Muñoz Martín, R. (2010): On Paradigms and Cognitive Translatology, in G. Schreve and E. Angelone (eds.) *Translation and Cognition*, Amsterdam and Philadelphia: John Benhamins, pp. 169–187.
- Mulder, G. (1979): Mental Load, Mental Effort and Attention. In N. Moray (Ed.), *Mental Workload: Its Theory and Measurement*. New York and London: Plenum Press.
- Nord, C. (2005): *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-oriented Text Analysis* (2nd Eds.). Amsterdam: Rodopi.
- O'Donnell, R. D. and Eggemeier, F. T. (1986): Workload Assessment Methodology. In: Boff, K. R., Kaufman, L. and Thomas, J. P. (eds) *Handbook of Perception and Human Performance, Vol. II: Cognitive Processes and Performance*. New York: Wiley. 42/41–42/49.
- PACTE. (2000): Acquiring Translation Competence: Hypotheses and Methodological Problems in a Research Project. In *Investigating Translation*, Beeby, A., Esinger, D., Presas, M. (eds). 99–106. Amsterdam: John Benjamins.
- PACTE. (2003): Building a Translation Competence Model. In: Alves, F. (ed.) *Triangulating Translation: Perspectives in Process Oriented Research*. Amsterdam: John Benjamins Pub, pp. 43–66.
- Plass, J. L., Moreno, R. and Brünken, R. (2010): Introduction. In: Plass, J. L., Moreno, R. and Brünken, R. (eds) *Cognitive Load Theory*. Cambridge; New York: Cambridge University Press, pp. 1–6.
- Pym, A. (2010): *Exploring Translation Theories*. London/New York: Routledge.
- RAND. (2002): Reading for understanding: toward a research and development program in reading comprehension Available from http://www.rand.org/pubs/monograph_reports/2005/MR1465.pdf.
- Rasinger, S. (2008): *Quantitative Research in Linguistics: An Introduction*, London: Continuum.
- Rehmann, A. J. (1995): *Handbook of Human Performance Measures and Crew Requirements for Flightdeck Research* (DOT/FAA/CT-TN95/49).
- Robinson, P. (ed.) (2011): *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*. Amsterdam; Philadelphia: John Benjamins.
- Rosenmund, A. (2001): Konstruktive Evaluation: Versuch Eines Evaluationskonzepts für den Unterricht. *Meta*, 46 (2), 301–310.

- Sabri, S. (2013): Item Analysis of Student Comprehensive Test for Research in Teaching Beginners Strings Ensemble Using Model Based Teaching Among Music Students in Public Universities, *International Journal of Education and Research*, 1 (12).
- Saldanha, G. and O'Brien, S. (2013): *Research Methodologies in Translation Studies*, Manchester, New York: St Jerome Publication.
- Sammer, G. (2006): Workload and Electro-encephalography Dynamics. In: Karwowski, W. (ed.) *International Encyclopedia of Ergonomics and Human Factors*. Boca Raton, FL: CRC/Taylor and Francis.
- Shreve, G. M., Danks, J. H. and Lacruz, I. (2004): *Cognitive Processes in Translation: Research Summary for the Center for the Advanced Study of Language, University of Maryland*.
- Shrock, S. A. and Coscarelli, W. C. C. (2007): *Criterion-referenced Test Development: Technical and Legal Guidelines for Corporate Training*. San Francisco: Pfeiffer.
- Spache, G. (1953): A New Readability Formula for Primary-Grade Reading Materials. *The Elementary School Journal*, 53 (7): 410–13. doi:10.1086/458513.
- Sun, S. (2015): Measuring Translation Difficulty: Theoretical and Methodological Considerations. *Across Language and Cultures*, 16 (1), 29–45.
- Taras, M. (2005): Assessment-Summative and Formative- Some Theoretical Reflections. *British Journal of Educational Studies*, 53 (4), 466–478.
- Vidulich, M.A. (1988): The Cognitive Psychology of Subjective Mental Workload. In: Hancock, P. A. and Meshkati, N. (eds) *Human Mental Workload*. Amsterdam/New York: North-Holland. 219–229.
- Vollmar, G. (2001): Maintaining Quality in the Flood of Translation Projects: A Model for Practical Quality Assurance. *The ATA Chronicle*, 30 (9), 24–27.
- Waddington, C. (2001): Different Methods of Evaluating Student Translations: The Question of Validity. *Meta : journal des traducteurs / Meta: Translators' Journal*, 46 (2), 311–325.
- Waddington, C. (2004): Should Student Translations be Assessed Holistically or through Error Analysis? *Lebende Sprachen*, 49 (1), 28–35.
- Wierwille, W. W., Eggemeier, F.T. (1993): Recommendations for Mental Workload Measurement in a Test and Evaluation Environment. *Human Factors*, 35 (2), 263–281.
- Wiersma, W. and Jurs, S. G. (1990): *Educational Measurement and Testing*. London: Allyn and Bacon.
- Williams, M. (2001): The Application of Argumentation Theory to Translation Quality Assessment. *Meta*, 46 (2), 326–344.
- Wilson, G. F. and Eggemeier, F. T. (2006): Mental Workload Measurement. In: Karwowski, W. (ed.) *International Encyclopedia of Ergonomics and Human Factors*. Boca Raton, FL: CRC/Taylor and Francis. 2nd ed., Vol. 1. 814–817.
- Wilss, W. (1982): *The Science of Translation: Problems and Methods*. Tübingen: G. Narr.
- Yeh, Y. Y., and Wickens, C. D. (1988): Dissociation of Performance and Subjective Measures of Workload. *Human Factors*, 30 (1), 111–120.