

Comparability of difficulty levels of translation tasks in CET-6 parallel test forms: evidence from product and process-based data

Yanmei Liu & Bingham Zheng

To cite this article: Yanmei Liu & Bingham Zheng (2022) Comparability of difficulty levels of translation tasks in CET-6 parallel test forms: evidence from product and process-based data, *The Interpreter and Translator Trainer*, 16:4, 428-447, DOI: [10.1080/1750399X.2022.2036938](https://doi.org/10.1080/1750399X.2022.2036938)

To link to this article: <https://doi.org/10.1080/1750399X.2022.2036938>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 14 Feb 2022.



[Submit your article to this journal](#)



Article views: 3712



[View related articles](#)



[View Crossmark data](#)





Citing articles: 1 [View citing articles](#)

ARTICLE

 OPEN ACCESS

 Check for updates

Comparability of difficulty levels of translation tasks in CET-6 parallel test forms: evidence from product and process-based data

Yanmei Liu ^a and Binghan Zheng ^b

^aSchool of Foreign Studies, Nanjing University of Posts and Telecommunications, Nanjing, China; ^bSchool of Modern Languages and Cultures, Durham University, Durham, UK

ABSTRACT

This study investigates the comparability of three parallel translation tasks selected from a College English Test Band-6 (CET-6) and explores the major linguistic features contributing to translation difficulty. Data obtained from the participants' subjective rating, eye-tracking, and performance evaluation were triangulated to measure the comparability of difficulty levels of parallel translation tasks. Data of word translation entropy, translation errors, and participants' retrospective reports were correlated to examine the difficulty triggers. The results show that: (i) the text comparability was evidenced by eye-tracking indicators and performance measurements, but not supported by subjective ratings; (ii) the domain content words (DCWs) were reported by the participants as the major cause of translation difficulties and the unequal number of DCWs among the three tasks led to inconsistent ratings for the task difficulty. Our findings suggest that test-takers' subjective perception and their cognitive skills deserve serious consideration by test designers, as these two factors can better demonstrate difficulty levels among parallel tasks. Our study postulates a new direction to establish a relationship between task characteristics and test validity, and provides suggestions for the CET-6 committee and other examination boards with practical methods to be able to compare the difficulty levels of parallel translation tasks.

ARTICLE HISTORY

Received 27 January 2021
Accepted 29 January 2022

KEYWORDS

Difficulty levels; Chinese-English translation; CET-6; domain content words; parallel test forms

1 Introduction

The College English Test Band-4 (CET-4) and Band-6 (CET-6) are administered twice a year by the National College English Testing Committee (NCETC) on behalf of the Higher Education Department in the Ministry of Education in China. Since its inception in 1987, the College English Test (CET) has attracted the Chinese public's attention as it has the largest number of test-takers in the world among all tests of English as a Foreign Language (EFL). The purpose of the CET is to examine the English proficiency of undergraduates in China and ensure that they have reached the required English levels specified in the National College English Teaching Syllabuses (NCETS) (Zheng and Cheng 2008). The examination consists of four sections: writing, listening, reading

CONTACT Binghan Zheng  binghan.zheng@durham.ac.uk

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

comprehension, and Chinese-English translation. In order to enhance test security, since 2012, the testing committee started using multiple parallel forms for each administration of the same test, with the writing and translation sections having three parallel forms.

Parallel forms refer to two or more testing versions that are interchangeable because they measure the same construct with the same purpose and are administered under the same conditions. They are widely used in large-scale standardised tests such as the International English Language Testing Service (IELTS), the Test of English as a Foreign Language (TOEFL), and the Graduate Record Examination (GRE). Adhering to the principle of test fairness, all test-takers should have the same opportunity to demonstrate their level of performance regardless of which form they are given (Bae and Lee 2011). A commonly held view is that the alternate forms should have inherent comparability in terms of testing items and difficulty levels. However, ‘examination boards are often criticised for their failure to provide evidence of comparability across forms, and few such studies are publicly available’ (Weir and Wu 2006, 167), which casts some doubt on the comparability of the difficulty level for the same testing item.

Existing research regarding task comparability has been conducted from both diachronic and synchronic perspectives. The diachronic approach centres upon task types in the same test over different periods, such as the independent and integrated writing tasks in the TOEFL test (Cumming et al. 2005; Plakans 2010). In contrast, the synchronic approach focuses either on crosswise comparisons between different tests with the same purpose, such as comparing the First Certificate in English (FCE) with the TOEFL test (Bachman et al. 1995; Kunnan and Carr 2017), or on different administration modes for the same test, such as comparing computer-based with paper-based TOEFL tests (Sawaki 2001; Choi, Kim, and Boo 2003). The comparability of multiple forms of the same test in one administration is also a major concern in this regard (Weir and Wu 2006; Li 2018). In relation to test fairness, the issue of task comparability is more critical for parallel forms of the same test in one administration than other types of parallel forms. Studies on this topic are scarce, with no research having been conducted concerning the comparability of three parallel translation tasks in the same CET-6 test. The present study, combining the product and process data, aims to fill this gap by addressing the following questions: (i) to what extent is the difficulty level of three parallel translation tasks comparable in a CET-6 test?; and (ii) what are the major factors leading to the disparity if the comparability of difficulty levels among translation tasks could not be achieved?

2 Literature review

2.1 Comparability of parallel test forms

Studies on the comparability of parallel test forms started in the 1980s when the Test of Written English (TWE) was administered with two different topics in a section of the TOEFL test. Stansfield and Ross (1988) argued that the comparability of scores obtained from the different topics should be carefully considered in order to secure the validity and reliability of the test. In addition, doubt concerning the comparability of parallel test forms used in the same administration or across different ones was also raised in the

1990s (Spolsky 1995; Bachman et al. 1995; Chalhoub-Deville and Turner 2000). In this section, we will review the relevant literature measuring task comparability across parallel forms of the same test, and the methods used for such studies.

Bachman, Davidson, and Milanovic (1996) analysed five experts' ratings on the characteristics of six parallel forms of FCE. Their results show that not all facets yielded substantive information about content comparability across the parallel forms. Based on the examinees' performances and the raters' judgements, Weir and Wu (2006) investigated the parallel forms reliability in the General English Proficiency Test Intermediate Speaking Test (GEPTS-I). The quantitative results show that all three GEPTS-I forms were parallel at the overall test level; but the qualitative analysis on raters' views of task difficulty revealed varied difficulty levels. Using the same methods as Weir and Wu (2006), Li (2018) analysed the comparability of picture-prompt writing tasks of three alternate CET-4 tests. The findings reveal that the performance data and the raters' subjective evaluations were generally consistent, but the raters expressed their reservations about the difficulty of the three parallel task prompts. Lei and Gu (2015) compared three parallel translation tasks by using 59 examinees' performance scores in a CET-4 test, and reported that there were possible discrepancies in the difficulty level among the three translation tasks. Although the above studies focused on different types of tasks in their investigations (writing, speaking, translation), the results showed that the parallel exam forms are not always comparable in their level of difficulty, which motivates us to investigate the comparability of parallel translation exam forms in a CET-6 test.

In terms of research methods, most existing studies used performance scores as the main or the only measurement for task difficulty. However, in an examination, test-takers will naturally invest greater cognitive effort on more challenging tasks in order to maintain the desired quality of their overall performance. Therefore, performance-based measurements cannot fully reveal the task difficulty, which calls for some additional evidence such as the test-taker's perception of the task difficulty and his/her cognitive efforts spent on a task. In the following section, the mixed-methods study on measuring translation difficulty by Campbell and Hale (1999), which was later applied in Sun (2015) and other recent studies, will be reviewed.

2.2 Measuring translation difficulty

Translation difficulty research, with its foci on what makes a text difficult to translate and how to measure the difficulty of a translation task (Sun 2019), has been advancing through interdisciplinary studies in the past two decades.

In terms of textual sources contributing to translation difficulty, some empirical studies have presented various types of linguistic items. Hale and Campbell (2002) indicated that a text with a high number of official terms, metaphors, and complex noun phrases could be regarded as a difficult text. Jensen (2009) made his initial attempt at proposing readability, word frequency, and non-literalness as major indicators of source text (ST) complexity in translation. Based on Jensen's indicators of translation text selection, Dragsted and Carl (2013) found that individual behavioural characteristics remained relatively constant across varying text complexity. Liu, Zheng, and Zhou (2019) further confirmed that the intrinsic complexity measured by readability, word frequency, and non-literalness was in line with the participants' subjective assessment of translation

difficulty. Using Support Vector Regression, Mishra, Bhattacharyya, and Carl (2013) claimed that translation difficulty can be predicted by three linguistic features: sentence length, degree of polysemy, and structural complexity.

Apart from the intrinsic difficulty indicated by the linguistic characteristics of ST, four types of external measurements including subjective rating, physiological measures, behavioural measures, and performance measures, have been frequently used in translation difficulty research. Subjective rating, be it a multi-dimensional scale that measures several specific aspects of cognitive load (Sun 2015), or a uni-dimensional scale that tests the overall cognitive load (Paas, Van Merriënboer, and Adam 1994), has been acknowledged as a valid, sensitive, and handy measurement of translation difficulty (Chen et al. 2016). Physiological and behavioural measures are mainly conducted using eye-tracking and key-logging methods, which allow for obtaining more precise determinations of cognitive effort (Lacruz 2017). O'Brien (2006) used eye-tracking data to identify different levels of cognitive effort when translators interacted with different fuzzy-match values presented by Translation Memory tools. Carl, Jakobsen, and Jensen (2008) suggested a triangulation model including fixation duration, fixation count, and pause data to investigate the cognitive effort of translators/interpreters, which has been applied to research on sight translation (Dragsted and Hansen 2009), written translation (Hvelplund 2011), and machine translation post-editing (Jia, Carl, and Wang 2019).

In addition, some researchers suggest that features of translation products can indicate translation difficulty to some extent. Campbell (1999) tested difficulty items by counting the number of alternative renditions in the target text (TT) produced by a group of subjects when translating the same ST. The author's results revealed that a source word will be rated as more difficult if there are more choices available for its translation. This idea was further developed into Choice Network Analysis (Campbell 2000), a framework which can model the mental processes of translation and estimate the difficulty level of STs. Carl, Schaeffer, and Bangalore (2016) termed the above framework as Word Translation Entropy (Htra), and used it to quantify the sum of all observed word translation options of a given ST word. Vanroy, De Clercq, and Macken (2019) correlated product features (number of errors, word translation entropy, and syntactic equivalence) with process features (duration, revision, and gaze information), and reported that translation difficulty reflected by process features can also be predicted by product features. This finding provides us with reasonable grounds for adopting Htra and error analysis to examine difficulty triggers during translation.

3 Methods

3.1 Participants

Thirty sophomores (mean age = 19.23, *SD* = 0.75 years) taking BA in Finance and BA in Business Administration programmes at Shandong University of Finance and Economics (China) participated in this research on a voluntary basis. They were all native Mandarin Chinese speakers with an average of 11.14 years of English (L2) learning (range = 11–12, *SD* = 0.35); had all passed CET-4 with a relatively high mean score at 584.86 (range = 558–627, *SD* = 20.23). However, they had not yet participated in a CET-6 test or taken any mock tests by themselves; and thus, were considered to be homogeneous in terms of

their English proficiency and familiarity with the topics of the experimental materials. They were all touch typists and had normal or corrected-to-normal vision; were asked to sign a consent form prior to the experiment, and rewarded with gift vouchers for their participation. The anonymity and confidentiality of the study were emphasised before the study began, and the experiment was approved by the research ethics committee of the university.

3.2 Source texts

The STs (see Appendix I) were three parallel translation tasks selected from a CET-6 administered in December 2017. The tasks related to the geography theme on individual lakes in China, and the length of the three source texts ranged from 158 to 176 Chinese characters, according to the computed results from *Chi-Editor*¹ (Bo et al. 2019). We used *Chi-Editor* to further test the STs in terms of the following aspects: the average sentence length, the longest sentence length, text difficulty, and the curriculum grade.² These textual features provided more information about the degree of content comparability across tasks. As can be seen from Table 1, although the three texts presented a slight variance in terms of text length, average sentence length, and the longest sentence length, they were all rated at a difficulty level of ‘High’ for the Curriculum Grade.

3.3 Experimental procedure

The participants were tested individually in the University’s eye-tracking lab. All the participants’ eye movements were registered using a Tobii T120 (120 Hz) eye-tracker attached to a 19-inch LCD monitor with a resolution of 1280 × 1024 pixels. The Chinese STs were displayed in the upper window of the Translog II user interface, with a typeface SimSun at 20-point size, and double line spacing. The English TTs were produced in the lower window, with the typeface New Times Roman at 20-point size, and double line spacing.

Prior to the formal experiments, all participants were asked to complete a short warm-up translation exercise. To avoid task-order effect, the formal experimental tasks were sequenced by Latin Square Design. Following the CET-6 instruction, the participants were asked to finish each translation task within 30 minutes without accessing any external resources. The total translation time for each participant being 1.5 hours; but a short break between tasks is allowed if requested by the participants.

After finishing all translation tasks, the participants were asked to rate the task difficulty on a 1–9 Likert scale, with 1 being extremely easy and 9 being extremely difficult. Subsequently, a questionnaire regarding difficulty triggers, together with the

Table 1. Textual features of the three STs.

ST	Text Length (character)	Average Sentence Length (character)	Longest Sentence Length (character)	Text Difficulty Level	Curriculum Grade
Text 1	167	23.86	36	3.60	High (6)
Text 2	176	29.33	42	3.72	High (6)
Text 3	158	22.57	42	3.28	High (5)

paper version of the three STs, was distributed to the participants for a retrospective interview. The following questions were used during interviews: a) At which level does the text pose difficulty for you in translation: lexical, syntactic, or textual?; b) Please highlight all translation difficulties in the texts; and c) For those translation difficulties, which process is more challenging for you: comprehension or expression?

3.4 Data processing and analysis

Quantitative data to ascertain translation difficulty were elicited concurrently from the subjective rating, eye-tracking, and performance measurements. Qualitative data to detect difficulty triggers were extracted from the participants' retrospective reports and translation error analysis using the Yet Another Word Alignment Tool (YAWAT tool) (Carl, Schaeffer, and Bangalore 2016).

3.4.1 Quantitative data analysis

Translation difficulty was measured from three dimensions: the participants' subjective perceptions of the cognitive load, the degree of their cognitive engagement in demanding tasks, and the performance evaluation by professional CET-6 raters.

A single item of a subjective scale, initially developed by Bratfisch, Borg, and Dornic (1972), had proven to be an effective way of assessing task difficulty by other practitioners as the indicator of overall cognitive load (e.g., Paas, Van Merriënboer, and Adam 1994). The present study adopted a uni-dimensional 1–9 Likert scale to reveal the participants' perceived cognitive load.

The engagement in the demanding translation tasks was recorded by the eye-tracker. Following Sharmin et al. (2008) and Sjørup (2013), fixation count and total fixation duration were applied to indicate cognitive effort allocated to the three parallel translation tasks. To ensure the quality of eye-tracking data, three rounds of screening were conducted.³ After that, eight out of the thirty participants were excluded from further analysis, with the percentage of invalid data being 26.67%. The log files were manually aligned using the YAWAT tool, through which the eye-tracking data were processed into a set of tables for analysis.

The final 66 translation products were assessed by three professional CET-6 raters who had over 10 years of experience in rating CET-6 translation tasks. The raters were asked to grade the translation products according to the rating criteria provided by the CET Committee (NCETC 2016, 10).

All the quantitative data were analysed by Linear Mixed Effects Regression (LMER) models in the R programme. With regard to the three dimensions of measurement data, we built four LMER models, taking the three STs as the fixed effect and the participants as the random effect. The dependent variables of these four LMER models were: 1) subjective rating score of task difficulty; 2) performance score; 3) total fixation duration; and 4) fixation count.

3.4.2 Qualitative data analysis

We applied four steps to identify translation difficulty: retrospective report, word-type tagging, translation error annotation, and correlation analysis. Firstly, the participants marked on the examination papers the points where they encountered difficulty during

translation. As the results show, the most difficult points were individual words rather than syntactic structures. We then tagged all Chinese words in the STs based on word segmentation using the YAWAT tool, into three categories: function word (FW), general content word (GCW), and domain content word (DCW).⁴ Thirdly, during the manual alignment of STs and TTs, translation errors were annotated by the researchers with types of addition/omission, minor mistranslation, and critical mistranslation using the YAWAT tool. Lastly, translation difficulty triggers were detected through the correlation analysis between translation errors, word translation entropy, and word types.

4 Results

4.1 Measurements of translation difficulty

4.1.1 Subjective rating

The mean score of subjective rating on translation difficulty shows a tendency of progressive increase, with Text 1 being the lowest while Text 3 is the highest (see Table 2). There was a significantly moderate agreement among the 22 participants that the level of translation difficulty across the three texts was not comparable (Kendall's $W = 0.615$, $p < .01$). The results from the LMER model, with the translation difficulty as the dependent variable, the texts as the fixed effect, and the participants as the random effect, reveal that the level of translation difficulty of Text 1 was significantly lower than that of Text 2 ($p = .000$) and Text 3 ($p = .000$) (see Figure 1).

4.1.2 Performance evaluation

To ensure the reliability of translation quality assessments, the inter-rater agreement was measured with Kendall's coefficient of concordance, with the W being 0.844, 0.718, and 0.919 respectively (see Table 3), indicating high internal consistency among the three raters.

The text effect on translation quality was conducted by another LMER model. The results (see Figure 2) show that there were no significant differences in terms of translation scores between Texts 1 and 2 ($p = .809$), Texts 1 and 3 ($p = .368$), and Texts 2 and 3 ($p = .256$), which implies that the level of translation difficulty was consistent among the three texts.

4.1.3 Eye-tracking measurements

Fixation count and total fixation duration were adopted as two eye-tracking indicators of cognitive effort the participants allocated to the tasks. As can be seen from Figure 3, from Text 1 to Text 3, fixation count shows an increasing tendency, with no significant differences between Texts 1 and 2 ($p = .322$), Texts 1 and 3 ($p = .058$), and Texts 2 and

Table 2. Statistical results of the subjective rating of translation difficulty.

Text	N	Mean	Sd.	Min	Max	Kendall's W	Chi-Square	Df	Sig
1	22	4.66	0.76	3.50	6.00	0.615	27.071	2	.000
2	22	5.59	0.92	4.50	7.00				
3	22	6.11	0.74	5.00	7.50				

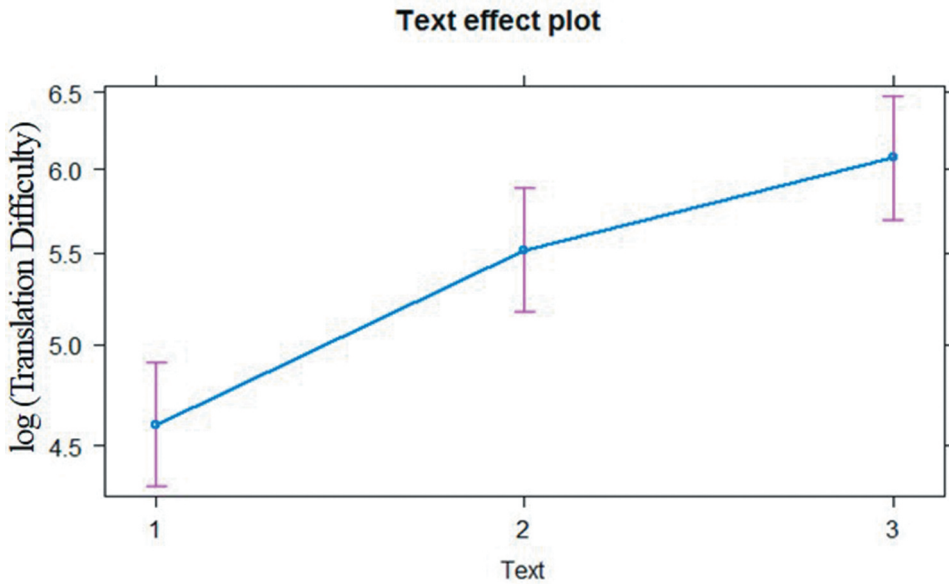


Figure 1. Plot of the effect of texts on participants' subjective rating of translation difficulty. (The dependent variables were logarithmically transformed to reduce skewness).

Table 3. Statistical results of the quality assessments of three translation tasks.

Text	N	Kendall's <i>W</i>	Chi-Square	Df	Sig.
1	3	0.844	53.144	21	.000
2	3	0.718	45.230	21	.002
3	3	0.919	57.881	21	.000

3 ($p = .217$); total fixation duration also shows an increasing tendency, with no significant differences between Texts 1 and 2 ($p = .843$), Texts 1 and 3 ($p = .502$), and Texts 2 and 3 ($p = .519$).

The translation quality assessment and eye-tracking data both indicate that irrespective of participants perceiving significantly different difficulties in the three translation tasks, there was no significant difference in the quality of their translations or the cognitive effort they invested in the tasks. The conflicting outcomes from the three dimensions of measurement made the comparability of the three texts complicated: on the one hand, participants thought that the degree of translation difficulty among the three texts was significantly different; on the other hand, judging from the effort they invested in the tasks and their performance results, the difficulty level of the three tests was indistinguishable.

Inconsistency between the results from subjective feedback and objective measurements has also been reported in previous studies on writing tests. According to Li (2018), the information abstractness of the task among the three test forms was the primary cause

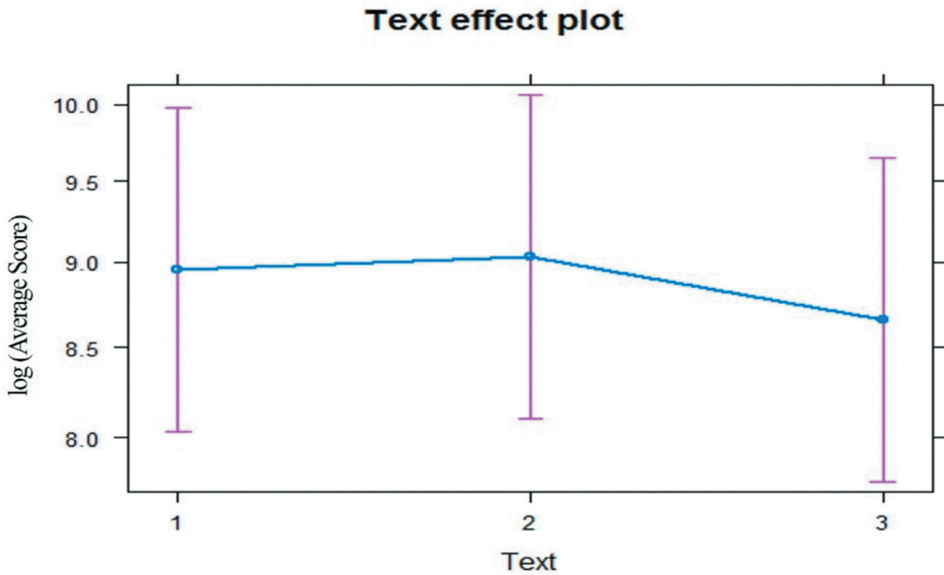


Figure 2. The effect of text on translation quality.

leading to these discrepancies. In addition, it is also plausible that there might be some difficulty triggers varying from task to task in the present study. Therefore, we investigated retrospective reports for the sources of difficulty in the translation tasks.

4.2 Translation difficulty triggers

4.2.1 Retrospective report

According to the retrospective data, the factors leading to translation difficulty mainly related to some specific words in the texts. Figure 4 illustrates difficult points in the three texts that were marked more than five times by all 22 participants.

It is clear that the number of high frequency difficult points (>5 times) in Text 3 was much higher than that in Text 1, and slightly higher than that in Text 2. This is in line with the participants' subjective assessment of translation difficulty, with Text 3 being the most difficult and Text 1 the least difficult.

Further examination found that most of the words that were marked as difficult can be categorised as domain-specific words (i.e., technical or jargon words), which are not commonly used in everyday language. According to Sung et al. (2016, 1245), 'every domain-specific word must have two values: domain specificity and conceptual difficulty.' It is thus speculated that these domain-specific words might be the sources of participants' translation difficulty. To prove this conjecture, we annotated translation errors to test the interactive effect between word categories and translation errors.

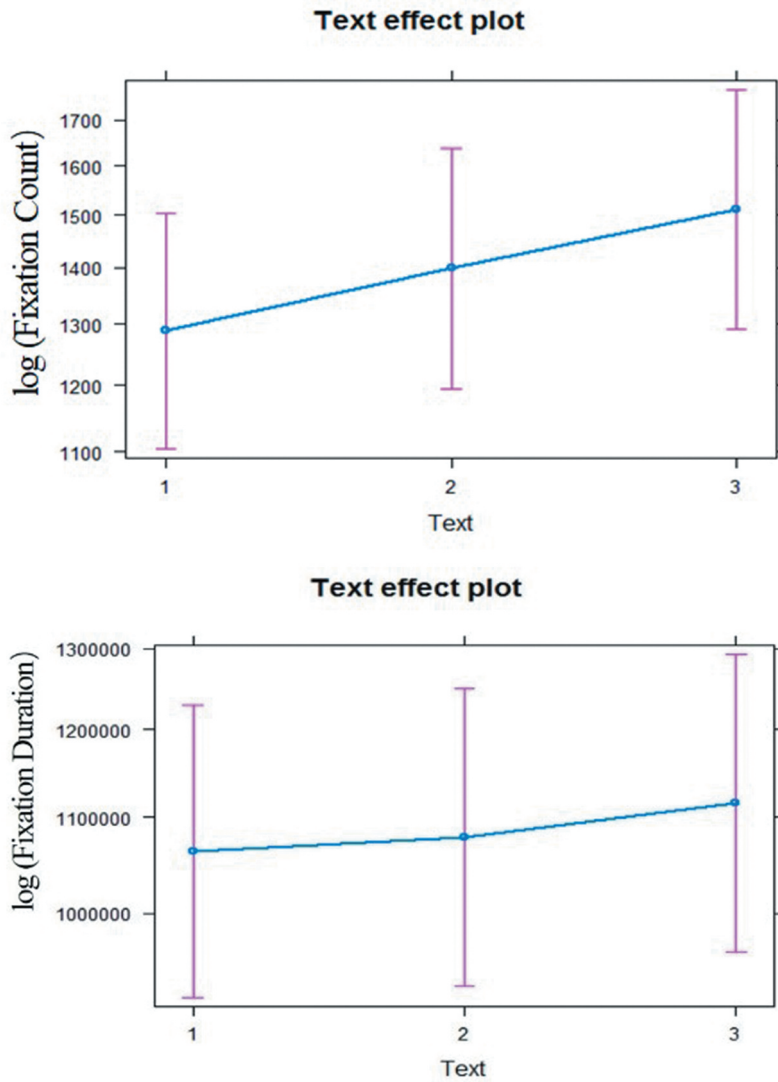


Figure 3. The effect of text on fixation count and total fixation duration.

4.2.2 Error analysis

This study used the YAWAT tool in which translation errors fall into three types: Addition/Omission,⁵ Mistranslation (Minor), and Mistranslation (Critical). According to the annotating principle of the YAWAT tool, the ST words should be aligned with the corresponding TT words as completely as possible.⁶ Any ST words for which no aligned TT words were found were annotated as 'Addition/Omission'. Words were deemed as 'Mistranslation' if the target content did not accurately represent the source content. Words which severely distorted source content were annotated as 'Critical Mistranslation', including wrong word choices, collocations, and predicate verbs.

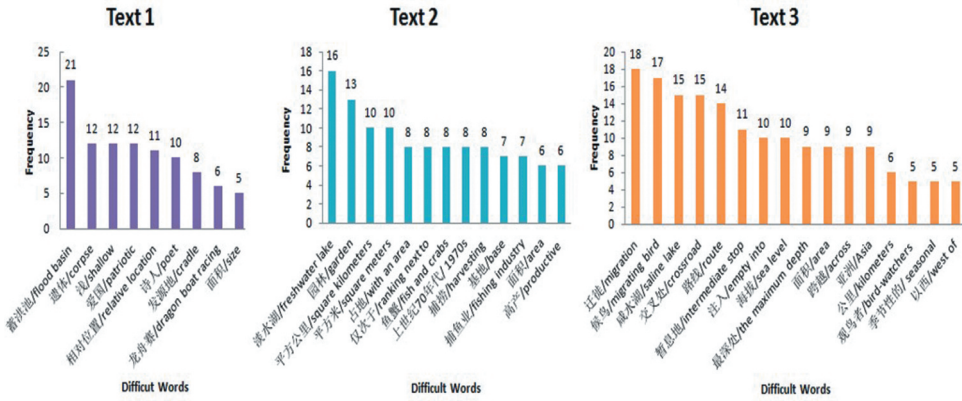


Figure 4. Translation difficult points marked in the three STs.

Problems such as spelling, punctuation, or redundancy were annotated as ‘Minor Mistranslations’. All the rest ST words which could be aligned with the corresponding TT words were categorised as ‘No Error’.

We assume that it was easier for difficult words to be mistranslated or expressed with additions or omissions. This phenomenon of word translation perplexity indicating ‘how many translation choices a translator has at a given point of the source text’ (Schaeffer et al. 2016, 29), namely, word translation entropy (Htra), describes the degree of uncertainty regarding which lexical TT item(s) is chosen, given the sample of alternative translations for a single ST word. Words with a higher Htra value indicate a greater level of perplexity, that is, they are more difficult to translate (Schaeffer et al. 2016).

Figure 5 shows the statistical Htra values for the four types of annotated alignments: Addition/Omission, Mistranslation (Minor), Mistranslation (Critical), and No Errors. Obviously, the Htra value of Mistranslation (Critical) was remarkably higher than that of

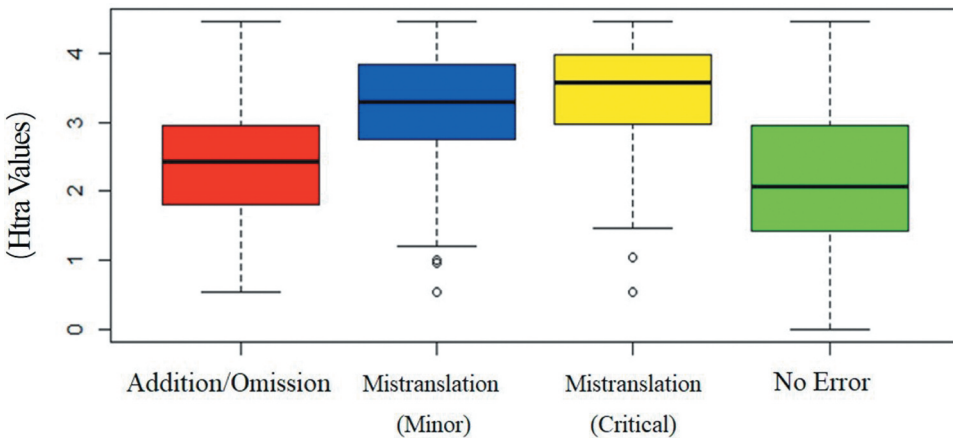


Figure 5. The interactive results of Htra values with four types of translation.

the other three categories. This indicates that the ST words annotated as Mistranslation (Critical) were the most difficult words to translate. The Htra value of Mistranslation (Minor) ranked second, Addition/Omission ranked third, and No Error ranked last.

As shown in Figure 5, translation difficulties were largely detected in the two types of mistranslation, especially in critical mistranslation. It is, therefore, necessary to further explore to what word categories these translation difficulties pertained to. In accordance with their functions in the sentence, all words in the ST were tagged as domain content word (DCW), general content word (GCW), or function word (FW). As shown in Figure 6, the correlation analysis of translation errors and word categories reveals that there was roughly the same proportion (about 10%) among the three-word categories for addition or omission. By contrast, mistranslations (minor and critical) took place mainly in DCWs, with the three types of translation errors reaching over 60% in DCWs, while remaining 15% and 20% in FWs and GCWs, respectively. In consideration of the higher Htra value of words with mistranslation errors, we can affirm that translation difficulties in these three tasks were attributed primarily to DCWs.

5 Discussion

Previous research on task difficulty of parallel test forms was largely based on participants' performance scores, whereas the discrepancies generated by different dimensions in the present study lends support to the argument that process data is a valuable addition to performance data (Paas et al. 2003). Performance scores of translation products may yield a biased result on task difficulty assessment because the interaction between performance and task difficulty or input load cannot be separated from the task operator's investment of cognitive effort. The relationship among these three aspects is presented in Figure 7.

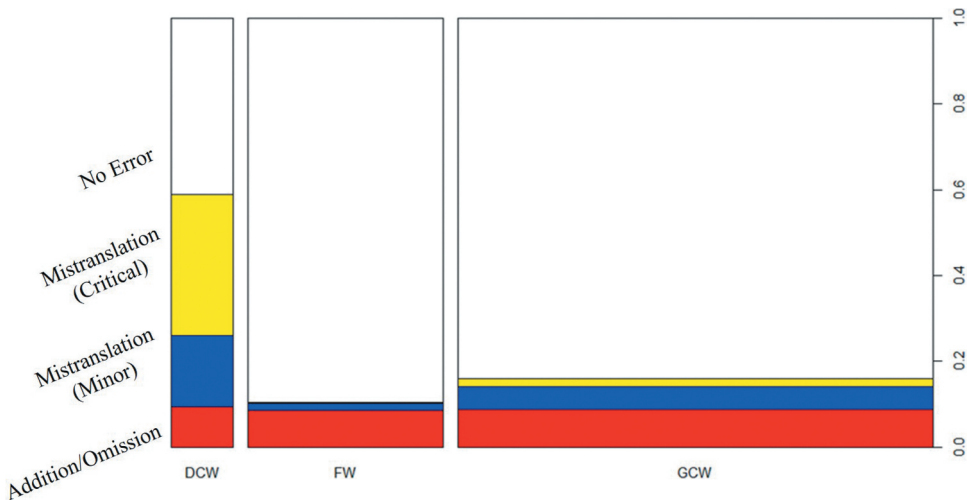


Figure 6. The interactive results of translation errors with different word categories.

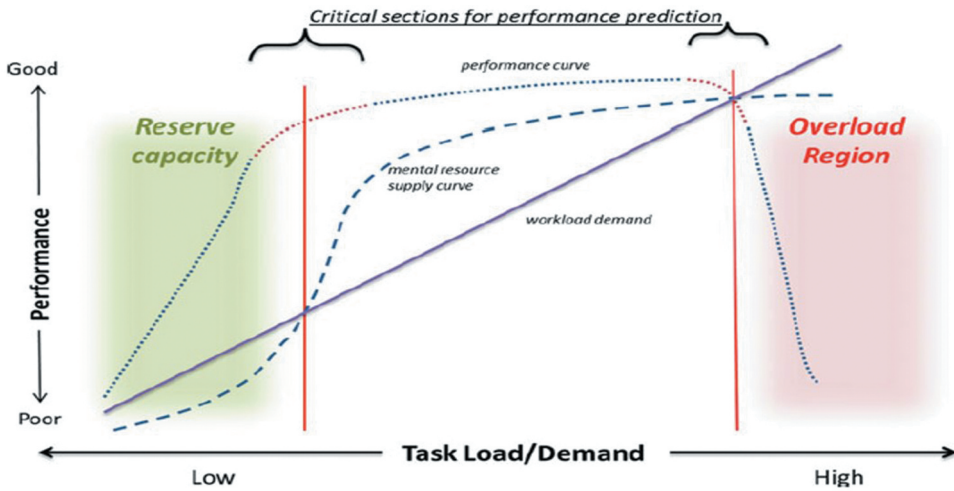


Figure 7. The relationship among performance, task load and cognitive effort (Chen et al. 2016, 39).

According to Johannsen (1979) and Chen et al. (2016), the task operator may be consciously ignoring (or unconsciously tolerating) some minor variations in workload demand. More mental resources are usually activated and supplied to meet the increase of mental demand (task complexity), leading to greater cognitive effort and maintaining the performance at a certain level. However, human working memory in information processing is limited in its capacity (see Sweller, Van Merriënboer, and Paas 1998; Wilson and Emmorey 2006). When the cognitive load reaches a certain threshold (the right redline of the workload as marked in Figure 7), further demand will lead to constant effort and further performance decrement characterised by an increase in errors (Chen et al. 2016). This cognitive overload is the principal contributor to the failure of task performance (Paas, Van Merriënboer, and Adam 1994). Similarly, Gile (1999) proposed a tightrope hypothesis in his effort models on simultaneous interpreting:

when the total capacity consumption is close to the interpreter's total available capacity, any increase in processing capacity requirements and any instance of mismanagement of cognitive resources by the interpreter can bring about overload or local attentional deficit and consequent deterioration of the interpreter's output (159).

Thus far, the reason for discrepancies among subjective rating, performance evaluation, and eye-tracking measurements regarding the task difficulty appears to be clear for the present study. The subjective rating on translation difficulty increases successively from Text 1 to Text 3 because the participants encountered the smallest number of DCWs in Text 1, but the greatest number in Text 3. In a translation examination setting when time was limited and consultation resources were not allowed, the participants were exposed to a situation where they could not allocate extra cognitive effort to translate DCWs, with their cognitive capacity reaching the overload region (cf. Figure 7). Relying only on their encyclopaedic knowledge, they had no capacity to resolve these translation problems, resulting in different degrees of mistakes. The retrospective reports from the participants also indicate that they tended to simplify or just skip the more insoluble translation problems. Therefore, their cognitive effort, indicated by eye-

tracking data, did not show significant differences among the three tasks. Furthermore, when the participants invested no extra cognitive resources in response to higher task demands, their performance shows no significant difference either, regardless of the slight decrease tendency in scores from Text 2 to Text 3 (see [Figure 2](#)).

What led to the tightrope point or the overload region in this study was DCWs, that were translated with more errors or omissions. A similar result was reported by Gile (1984), that there was a higher rate of failure in rendering proper names in simultaneous interpreting. Hale and Campbell (2002) also identified official terms as one of the key translation difficulties. In contrast to difficulty indices proposed in previous studies (Hale and Campbell 2002; Jensen 2009; Mishra, Bhattacharyya, and Carl 2013), DCWs serve as the decisive factor of difficulty triggers in the present study. The possible reasons for the distinction are as follows: on the one hand, the translation direction (L1 to L2) leaves little trouble in ST comprehension for the participants; on the other hand, as less experienced translators, the participants' 'cognitive operations tend to be bottom-up' (Zheng 2012, 169), giving their priority to smaller units such as lexical problems in translation. Lörcher (1991, 1993, 2005) found that foreign language students, compared to professional translators, concentrated especially on single words, paying little attention to stylistic and text-type adequacy. Barbosa and Neiva (2003) also reported that less experienced translators employed smaller units such as single words, phrases, or clauses: 'No units were as long as a sentence or a paragraph' (139).

Our finding about DCWs leading to the participants' cognitive overload reveals that the examinees' perceived translation difficulty was predominantly caused by this category of words, rather than by sentence structure and textual cohesion, which remain major concerns for the CET Committee. According to the guideline of the CET programme (NCETC 2016, 4), examinees are required to translate a passage about familiar subjects at a medium difficulty level accurately and smoothly from Chinese to English. To assess the examinees' translation competence, special emphasis has been placed on the following three aspects: message transforming at sentential and textual levels, and the proper use of translation strategy. From the annual tests of parallel forms of passage translation in CET-6 from June 2013 to September 2020 (with 45 parallel test papers in total), all of the translation items have distinct domain subjects, such as architecture, economics, transportation, geography, Chinese dynasties, and Chinese classical novels. DCWs have played a vital role in dominating topic familiarity and posing translation difficulty, thereupon influencing examinees' translation behaviours and performance. In this regard, it is particularly important to balance the amount of DCWs and the degree of their specialisation in a specific domain, in order to establish a comparison of difficulty levels in translation tasks with distinct domain subjects. As a further guarantee of test validity, a pre-test is essential to obtain feedback from quasi test-takers on the translation difficulty level and difficulty triggers.

6 Conclusion

Driven by curiosity about the comparability of translation tasks in the CET-6 parallel forms, this study investigated the degree of translation difficulty from three dimensions: the participants' subjective perceptions of the cognitive load, the degree of their cognitive engagement in demanding tasks, and the performance evaluation by professional raters.

Given that the results from the three dimensions point to different directions, we endeavoured to find the reasons for this by analysing the linguistic features and the participants' retrospection data. Suggestions based on our results are as follows: Firstly, although the three texts are comparable according to the results of performance evaluation and eye-tracking measures, participants' subjective rating results reveal that the parallel translation tasks are not of equal difficulty. Secondly, DCWs have proved to be the major cause of translation difficulty as perceived by the participants, which results in more errors or omissions concerning the difficulty triggers.

Taking all the results into consideration, we argue that multi-dimensional measurement is essential to improve the comparability and reliability of parallel testing forms. Apart from the performance scores, the examinees' subjective perception and cognitive efforts on tasks are helpful complements to control the difficulty levels of parallel tasks. Content analytic data are equally inspiring for test designers to provide reassurance that the designed tests are appropriate for potential examinees. In addition to the factors manipulating task difficulty such as the topics, text types, grammatical structures, and lexical familiarity suggested in writing and speaking tests (Weir and Wu 2006; Li 2018), the present research demonstrates the possibility that lexical factors could outweigh the other text characteristics contributing to task difficulty. By correlating Htra value and translation errors, we found that the translation of DCWs was the major challenge for participants in the examination setting and consumed a large proportion of their cognitive efforts. This finding provides practical measures for test designers in preparing parallel translation tests for languages learners. In particular, considering the conventions of distinct domain subjects in CET translation tasks as well as the various academic backgrounds of the potential examinees, deliberate control over the degree of specialisation from the perspective of translation rather than reading comprehension deserves more attention from test designers to meet the requirements of 'translating familiar subjects of Chinese with medium-difficulty into English' made by CET programme (NCETC 2016, 4).

The present study is limited by the small number of tasks, which may not be sufficient to account for a full picture of translation testing in CET-6. Besides, CET-6 is limited to Chinese-to-English translation direction, so whether the results could apply to the opposite direction in other tests need to be verified in further studies. Our explorative results demonstrate the potential influence of the amount of DCWs on translation difficulty, and this finding invites further research into the possible impacts of other qualities of DCWs (e.g., word frequency, part of speech) on test-takers' cognitive effort and performance. Furthermore, translation difficulty research should involve at least three factors: the test-takers, the text, and the raters. This study gives much weight to the former two factors, leaving the raters' perception and rating process less addressed. It is thus essential to incorporate the raters' role into our future research.

Notes

1. Chi-Editor is an online text evaluation and adaptation system that matches Chinese reading texts to specific proficiency levels specified in the International Curriculum for Chinese Language Education (Confucius Institute Headquarters 2015), which annotates texts with a

number of lexical and syntactic features to inform text adaptation. The system references vocabulary lists from national Chinese as second language (CSL) curriculum standards. A corpus of approximately 550 widely-used CSL textbooks is used to provide benchmarks for text complexity evaluation and lexical and syntactic annotation.

2. According to Chi-Editor, the curriculum grade, based on the standard of *International curriculum for Chinese language education*, ranges from 'One' which is the easiest, to 'Six', the most difficult. Grades One and Two belong to the primary level, corresponding in text difficulty to 1-1.5 and 1.5-2.0, respectively; The intermediate level includes Grades Three and Four, corresponding in text difficulty to 2.0-2.5 and 2.5-3.0, respectively; and the high level consists of Grades Five and Six, corresponding in text difficulty to 3.0-3.5 and 3.5-4.0, respectively.
3. Based on the reports from Rayner (1998), fixations usually last 200-300 ms. Any mean fixation duration lower than 200 ms in the present study was thus filtered out from analysis. Following Sjørup's (2013) criterion, we eliminated the recordings with an average GTS (gaze time on screen) lower than 65%. The remaining data were double-checked from Translog files, with some outlier keystroke data being further removed.
4. Domain content words, one of the linguistic features implemented in the Chinese Readability Index Explorer (CRIE), refer to content words with domain knowledge (Sung et al. 2016). General content words refer to all other content words except domain content words.
5. Addition or omission is marked by a single error type (Addition/Omission) in the YAWAT tool. The source and target text alignments have been manually annotated for the error types. Addition/Omission is categorised as a type of errors because we define 'Addition' as 'over-translated items' and 'Omission' as 'under-translated items' appearing in the target text. They do not include strategic explicitation and implicitation of the ST meaning, which are aligned as 'No Error'.
6. The alignment between the ST and the TT was manually operated by the researchers based on dynamic meaning equivalence. An aligned ST word in this study refers to a minimal meaningful unit which may contain one or more Chinese characters.

Acknowledgments

The authors wish to thank all the participants who took part in the experiment and Miss Tingting Jia for helping with data collection. We are grateful to Dr. Han Chao and two anonymous reviewers for their invaluable comments which helped us improve this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Social Science Fund of China [No. 19BY124].

ORCID

Yanmei Liu  <http://orcid.org/0000-0001-8050-7554>

Binghan Zheng  <http://orcid.org/0000-0001-5302-4709>

References

- Bachman, L. F., F. Davidson, and M. Milanovic. 1996. "The Use of Test Method Characteristics in the Content Analysis and Design of EFL Proficiency Tests." *Language Testing* 13 (2): 125–150. doi:10.1177/026553229601300201.
- Bachman, L. F., F. Davidson, K. Ryan, and I.-C. Choi. 1995. *An Investigation of the Comparability of the Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*. Cambridge: Cambridge University Press.
- Bae, J., and Y. S. Lee. 2011. "The Validation of Parallel Test Forms: 'Mountain' and 'Beach' Picture Series for Assessment of Language Skills." *Language Testing* 28 (2): 155–177. doi:10.1177/0265532210382446.
- Barbosa, H. G., and A. M. S. Neiva. 2003. "Using Think-aloud Protocols to Investigate the Translation Process of Foreign Language Learners and Experienced Translators." In *Triangulating Translation: Perspectives in Process Oriented Research*, edited by F. Alves, 137–156. Amsterdam: John Benjamins Publishing.
- Bo, W., J. Chen, K. Guo, and T. Jin. 2019. "Data-driven Adapting for Fine-tuning Chinese Teaching Materials: Using Corpora as Benchmarks." In *Computational and Corpus Approaches to Chinese Language Learning*. *Chinese Language Learning Sciences*, edited by X. Lu and B. Chen, 99–118. Singapore: Springer. doi:10.1007/978-981-13-3570-9_6.
- Bratfisch, O., G. Borg, and S. Dornic. 1972. "Perceived Item-difficulty in Three Tests of Intellectual Performance Capacity." Reports from Institute of Applied Psychology, the University of Stockholm. <https://files.eric.ed.gov/fulltext/ED080552.pdf>
- Campbell, S. 1999. "A Cognitive Approach to Source Text Difficulty in Translation." *Target* 11 (1): 33–63. doi:10.1075/target.11.1.03cam.
- Campbell, S. 2000. "Choice Network Analysis in Translation Research." In *Intercultural Faultlines: Research Models in Translation Studies: Textual and Cognitive Aspects*, edited by M. Olohan, 29–42. Manchester: St. Jerome.
- Campbell, S., and S. Hale. 1999. "What Makes a Text Difficult to Translate." *Refereed Proceedings of the 23rd Annual ALAA Congress*. <http://www.cltr.uq.edu.au/alaa/proceed/camp/hale.html>
- Carl, M., A. L. Jakobsen, and K. T. H. Jensen. 2008. "Studying Human Translation Behavior with User Activity Data." In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science*, edited by B. Sharp and M. Zock, 114–123. Barcelona: Insticc Press.
- Carl, M., M. Schaeffer, and S. Bangalore. 2016. "The CRITT Translation Process Research Database." In *New Directions in Empirical Translation Process Research*, edited by M. Carl, S. Bangalore, and M. Schaeffer, 13–54. New York: Springer. doi:10.1007/978-3-319-20358-4_2.
- Chalhoub-Deville, M., and C. E. Turner. 2000. "What to Look for in ESL Admission Tests: Cambridge Certificate Exams, IELTS, and TOEFL." *System* 28 (4): 523–539. doi:10.1016/S0346-251X(00)00036-1.
- Chen, F., J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji, and D. Conway. 2016. *Robust Multimodal Cognitive Load Measurement*. Cham: Springer.
- Choi, I. C., K. S. Kim, and J. Boo. 2003. "Comparability of a Paper-based Language Test and a Computer-based Language Test." *Language Testing* 20 (3): 295–320. doi:10.1191/0265532203lt258oa.
- Confucius Institute Headquarters. 2015. *International Curriculum for Chinese Language Education*. Beijing: Beijing Language and University Press.
- Cumming, A., R. Kantor, K. Baba, U. Erdosy, K. Eouanzoui, and M. James. 2005. "Differences in Written Discourse in Independent and Integrated Prototype Tasks for Next Generation TOEFL." *Assessing Writing* 10 (1): 5–43. doi:10.1016/j.asw.2005.02.001.
- Dragsted, B., and M. Carl. 2013. "Towards a Classification of Translation Styles Based on Eye-tracking and Keylogging Data." *Journal of Writing Research* 5 (1): 133–158. doi:10.17239/jowr-2013.05.01.6.

- Dragsted, B., and I. G. Hansen. 2009. "Exploring Translation and Interpreting Hybrids. The Case of Sight Translation." *Meta* 54 (3): 588–604. doi:10.7202/038317ar.
- Gile, D. 1984. "Les Noms Propres En Interprétation Simultanée." *Multilingua-Journal of Cross-cultural and Interlanguage.* *Communication* 3 (2): 79–86.
- Gile, D. 1999. "Testing the Effort Models' Tightrope Hypothesis in Simultaneous Interpreting-A Contribution." *HERMES-Journal of Language and Communication in Business* 23: 153–172. doi:10.7146/hjlc.v12i23.25553.
- Hale, S., and S. Campbell. 2002. "The Interaction between Text Difficulty and Translation Accuracy." *Babel* 48 (1): 14–33. doi:10.1075/babel.48.1.02hal.
- Hvelplund, K. T. 2011. *Allocation of Cognitive Resources in Translation: An Eye-Tracking and Key-Logging Study*. Ph.D. thesis. Copenhagen: Copenhagen Business School. <http://hdl.handle.net/10419/208778>
- Jensen, K. T. 2009. "Indicators of Text Complexity." In *Behind the Mind: Methods, Models and Results in Translation Process Research*, edited by A. L. Jakobsen and I. M. Mees, 61–80. Copenhagen: Samfundslitteratur.
- Jia, Y., M. Carl, and X. Wang. 2019. "How Does the Post-editing of Neural Machine Translation Compare with From-scratch Translation: A Product and Process Study." *The Journal of Specialised Translation* 31: 60–86.
- Johannsen, G. 1979. "Workload and Workload Measurement." In *Mental Workload: Its Theory and Measurement*, edited by N. Moray, 3–11. New York: Plenum Press. doi:10.1007/978-1-4757-0884-4_1.
- Kunnan, A. J., and N. T. Carr. 2017. "A Comparability Study between the General English Proficiency Test-Advanced and the Internet-Based Test of English as A Foreign Language." *Language Testing in Asia* 7 (1): 7–17. doi:10.1186/s40468-017-0048-x.
- Lacruz, I. 2017. "Cognitive Effort in Translation, Editing, and Post-editing." In *The Handbook of Translation and Cognition*, edited by J. Schwieter and A. Ferreira, 386–401. Malden, MA: John Wiley & Sons.
- Lei, X., and X. Gu. 2015. "Are CET-4 Parallel Translation Tasks at the Same Difficult Level? -A Case Study of CET-4 Translation Tasks in June 2014." *Foreign Language Testing and Teaching*, 17 (1): 18–23.
- Li, J. 2018. "Establishing Comparability across Writing Tasks with Picture Prompts of Three Alternate Tests." *Language Assessment Quarterly* 15 (4): 368–386. doi:10.1080/15434303.2017.1405422.
- Liu, Y., B. Zheng, and H. Zhou. 2019. "Measuring the Difficulty of Text Translation: The Combination of Text-focused and Translator-oriented Approaches." *Target* 31 (1): 125–149. doi:10.1075/target.18036.zhe.
- Lörscher, W. 1991. *Translation Performance, Translation Process, and Translation Strategies: A Psycholinguistic Investigation*. Tübingen: Gunter Narr.
- Lörscher, W. 1993. "Translation Process Analysis." In *Translation and Knowledge*, edited by Y. Gambier and J. Tömmola, 195–212. Turku: University of Turku, Center for Translation and Interpreting.
- Lörscher, W. 2005. "The Translation Process: Methods and Problems of Its Investigation." *Meta* 50 (2): 597–608. doi:10.7202/011003ar.
- Mishra, A., P. Bhattacharyya, and M. Carl. 2013. "Automatically Predicting Sentence Translation Difficulty." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short papers)*, edited by H. Schuetze, P. Fung, and M. Poesio, 346–351. Sofia: Association for Computational Linguistics.
- NCETC (National College English Testing Committee). 2016. *The Guideline for College English Test (CET-4 and CET-6)*. <http://cet.neea.edu.cn/html1/folder/16113/1588-1.htm>
- O'Brien, S. 2006. "Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output." *Across Languages and Cultures* 7 (1): 1–21. doi:10.1556/Acr.7.2006.1.1.
- Paas, F. G., J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven. 2003. "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory." *Educational Psychologist* 38 (1): 63–71. doi:10.1556/Acr.7.2006.1.1.

- Paas, F. G., J. J. Van Merriënboer, and J. J. Adam. 1994. "Measurement of Cognitive Load in Instructional Research." *Perceptual and Motor Skills* 79 (1): 419–430. doi:10.2466/pms.1994.79.1.419.
- Plakans, L. 2010. "Independent Vs. Integrated Writing Tasks: A Comparison of Task Representation." *TESOL Quarterly* 44 (1): 185–194. doi:10.5054/tq.2010.215251.
- Rayner, K. 1998. "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological Bulletin* 124 (3): 372–422. <https://psycnet.apa.org/doi/10.1037/0033-2909.124.3.372>
- Sawaki, Y. 2001. "Comparability of Conventional and Computerized Tests of Reading in a Second Language." *Language Learning & Technology* 5 (2): 38–59.
- Schaeffer, M., B. Dragsted, K. T. Hvelplund, L. W. Balling, and M. Carl. 2016. "Word Translation Entropy: Evidence of Early Target Language Activation during Reading for Translation." In *New Directions in Empirical Translation Process Research*, edited by M. Carl, S. Bangalore, and M. Schaeffer, 181–210. New York: Springer.
- Sharmin, S., O. Špakov, K. J. Rähä, and A. L. Jakobsen. 2008. "Where on the Screen Do Translation Students Look while Translating, and for How Long?" In *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*, edited by S. Göpferich, A. L. Jakobsen, and I. M. Mees, 31–51. Copenhagen: Samfundslitteratur.
- Sjørup, A. C. 2013. *Cognitive Effort in Metaphor Translation: An Eye-Tracking and Key-Logging Study*. Ph.D. thesis. Copenhagen: Copenhagen Business School. <http://hdl.handle.net/10419/208853>
- Spolsky, B. 1995. *Measured Words: The Development of Objective Language Testing*. Oxford: Oxford University Press.
- Stansfield, C. W., and J. Ross. 1988. "A Long-term Research Agenda for the Test of Written English." *Language Testing* 5 (2): 160–186. doi:10.1177/026553228800500204.
- Sun, S. 2015. "Measuring Translation Difficulty: Theoretical and Methodological Considerations." *Across Languages and Cultures* 16 (1): 29–54. doi:10.1556/084.2015.16.1.2.
- Sun, S. 2019. "Measuring Difficulty in Translation and Post-editing: A Review." In *Researching Cognitive Processes of Translation*, edited by D. Li, V. Lei, and Y. He, 139–168. Singapore: Springer. doi:10.1007/978-981-13-1984-6_7.
- Sung, Y., T. Chang, W. Lin, K. Hsieh, and K. Chang. 2016. "CRIE: An Automated Analyzer for Chinese Texts." *Behavior Research Methods* 48: 1238–1251. doi:10.3758/s13428-015-0649-1.
- Sweller, J., J. J. Van Merriënboer, and F. G. Paas. 1998. "Cognitive Architecture and Instructional Design." *Educational Psychology Review* 10 (3): 251–296. doi:10.1023/A:1022193728205.
- Vanroy, B., O. De Clercq, and L. Macken. 2019. "Correlating Process and Product Data to Get an Insight into Translation Difficulty." *Perspectives* 27 (6): 924–941. doi:10.1080/0907676X.2019.1594319.
- Weir, C. J., and J. R. W. Wu. 2006. "Establishing Test Form and Individual Task Comparability: A Case Study of A Semi-direct Speaking Test." *Language Testing* 23 (2): 167–197. doi:10.1191/0265532206lt326oa.
- Wilson, M., and K. Emmorey. 2006. "Comparing Sign Language and Speech Reveals a Universal Limit on Short-term Memory Capacity." *Psychological Science* 17 (8): 682–683. doi:10.1111/j.1467-9280.2006.01766.x.
- Zheng, B. 2012. *Choice-making in the Process of English-Chinese Translation: An Empirical Study*. Beijing: Foreign Language Teaching and Research Press. doi:10.32657/10356/14501.
- Zheng, Y., and L. Cheng. 2008. "Test Review: College English Test (CET) in China." *Language Testing* 25 (3): 408–417. doi:10.1177/0265532208092433.

Appendix: Source Texts (CET-6, Dec 2017)

Text 1: 洞庭湖位于湖南省东北部, 面积很大, 但湖水很浅。洞庭湖是长江的蓄洪池, 湖的大小很大程度上取决于季节变化。湖北和湖南两省因其与湖的相对位置而得名, 湖北意为湖的北边, 而湖南则为湖的南边。洞庭湖作为龙舟赛的发源地, 在中国文化中享有盛名。据说龙舟赛始于洞庭湖东岸。为的是搜寻楚国爱国诗人屈原的遗体。龙舟赛与洞庭湖及周边的美景, 每年都吸引着成千上万来自全国和世界各地的游客。

Text 2: 太湖是中国东部的一个淡水湖, 占地面积 2250 平方公里, 是中国第三大淡水湖, 仅次于鄱阳和洞庭。太湖约有 90 个岛屿, 大小从几平方米到几平方公里不等。太湖以其独特的太湖石而闻名, 太湖石常用于装饰中国传统园林。太湖也以高产的捕鱼业闻名。自上世纪 70 年代后期以来, 捕捞鱼蟹对沿湖的居民来说极为重要, 并对周边地区的经济作出了重大贡献。太湖地区是中国陶瓷业基地之一, 其中宜兴的陶瓷(ceramics)厂家生产举世闻名的宜兴紫砂壶(clay teapot)。

Text 3: 青海湖位于海拔 3205 米, 青海省省会西宁以西约 100 公里处, 是中国最大的咸水湖, 面积 4317 平方公里, 最深处 25.5 米。有 23 条河注入湖中, 其中大部分是季节性的。百分之八十的湖水源于五条主要河流。青海湖位于跨越亚洲的几条候鸟迁徙路线的交叉处。许多鸟类把青海湖作为迁徙过程中的暂息地, 湖的西侧是著名的鸟岛, 吸引着来自世界各地的观鸟者。每年夏天, 游客们也来这里观看国际自行车比赛。