

## Effects of thinking aloud on cognitive effort in translation

### Sanjun Sun

Beijing Foreign Studies University, China  
sunsanjun@bfsu.edu.cn  
<https://orcid.org/0000-0002-4020-8418>

### Tian Li

Beijing Foreign Studies University, China  
litian@bfsu.edu.cn  
<https://orcid.org/0000-0001-7320-204X>

### Xiaoyan Zhou

Beihang University, China  
colleenzhou@126.com  
<https://orcid.org/0000-0002-8965-5585>

### Abstract

This study investigates the effects of thinking aloud on cognitive effort in translation as a function of source-text difficulty level. It does so by considering time on task, duration of different translation phases (i.e., orientation, draft, revision), cognitive effort of processing the source and target texts, and translation quality. Twenty participants took part in an English–Chinese translation experiment, which comprised two matched sessions – translating while thinking aloud and translating silently. Their translation processes were recorded by means of an eye tracker and a key logger. An adapted NASA Task Load Index was employed to elicit their subjective assessments of translation difficulty levels. The quality of their translations was evaluated. The results of the study reveal a number of important effects of thinking aloud on cognitive effort during translation: on translation duration, cognitive effort, the perceived level of difficulty of a translation as measured by NASA-TLX, and on translating easier texts.

**Keywords:** think-aloud protocols; translation; cognitive effort; eye-tracking; key-logging

## 1. Introduction

As a data elicitation method, thinking aloud (TA) refers to verbalizing one's thoughts out loud while engaging in a task, and written transcripts of the verbalizations are called think-aloud protocols (TAPs) (Ericsson & Simon, 1993; Jääskeläinen, 2010). In translation process research, it used to be a primary method for investigating translators' cognitive processes from the mid-1980s until the end of the 1990s. In recent years, TA has been adopted by some translation researchers (e.g., Angelone, 2010; Vieira, 2017). And yet, for various reasons, TAP-based translation studies have dwindled noticeably. One reason is the emerging and increasing use of keystroke logging, eye-tracking and other methods (see Jakobsen, 2017). Another reason is concerns and doubts about TA's validity and comprehensiveness (Hansen, 2005; Jääskeläinen, 2011). To date, few studies (e.g., Jakobsen, 2003) have empirically tested TA's validity, despite calls for more research into this method (Jääskeläinen, 2017).

This study is an empirical investigation of the impact of TA on the translation process. It seeks to examine the effects of TA on cognitive effort in translation by considering time on task, duration of different translation phases, cognitive effort in processing the source texts (STs) and target texts (TTs), and translation quality, as functions of the difficulty of STs. It bases its analysis on verbal, keystroke logging and eye-tracking data collected from the experiment, and post-performance rating of the level of difficulty of a translation. The methods of data collection are complementary and they aim to provide an accurate picture of the effects of TA on the translation process.

We first review and discuss the theories, arguments and recent empirical findings concerning TAPs. Then we introduce the concept of *cognitive effort*, its position in translation process research and its various measures before we describe our experiment set-up and present the findings.

## 2. Validity of TAPs

The theory that verbal protocols can be used to elicit data on cognitive processes was proposed by Ericsson and Simon (1980, 1993), who have provided substantial empirical support for it. Ericsson and Simon (1993) hold that

“subjects can generate verbalizations, subordinated to task-driven cognitive processes (think aloud), without changing the sequence of their thoughts, and slowing down only moderately due to the additional verbalization” (p. xxxii).

Ericsson and Simon (1993) state that TAPs provide “a nearly complete record of the sequence of information that is heeded during task performance” (p. 62); in contrast, processes that are

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

automated or unconscious are unavailable for TA. The controversies concerning the validity of TAPs go around these two points: (1) whether TA alters the thought processes being studied (reactivity), and (2) whether TAPs are an accurate reflection of thoughts (veridicality; Bowles, 2018). Recently, Fox et al. (2011), in a meta-analysis of 94 studies (involving almost 3,500 participants) comparing performance while TA to that in a matching silent condition, found that TA results in little or no reliable difference in performance across TA and silent conditions, though it does prolong the time taken to reach solution.

Most of the studies included in the meta-analysis of Fox et al. (2011) involved problem-solving tasks (e.g., playing chess), making the relevance of their findings to linguistic tasks (such as translation and reading) unclear (Bowles, 2018). Partly for this reason, since the early years of translation process research, the validity and completeness of TAPs have been controversial. Toury (1991), for example, suspected that “spoken” translation (i.e., TA) and “written” translation might involve different strategies and therefore TA might interfere with translation. Hansen (2005) contended that TA “must have an impact ... on the thought processes, on the translation process and on the translation product” (p. 519). Sun (2011), in his review, analysed these claims and argued that no strong evidence suggests that TA significantly changes or influences the translation process, although several variables (e.g., task or text characteristics) in a specific study might influence the validity and completeness of TAPs.

Two oft-cited empirical studies in this regard are Krings (2001) and Jakobsen (2003). Krings (2001), in his study of the process of post-editing machine-translated texts as well as translation, found that TA slowed down the process by roughly 30%. Jakobsen (2003) reported that TA delayed translation by about 25% and forced translators to process text in smaller segments. Their findings have been interpreted as evidence that TA might change the course or structure of cognitive processes (e.g., Nitzke, 2019, p. 92). This, however, is a misunderstanding of their findings and the expression “course or structure of cognitive processes”.

Metaphorically speaking, if a person follows the same path as usual, we would say their path (or course) is unchanged, even when they walk more slowly and take smaller steps than usual. Cognitive structure means “rules for processing information or for connecting experienced events” (Kohlberg, 1969, p. 349) and it is exhibited in four types of statement (Ericsson & Simon, 1993, p. 198):

- *intentions* (goals and future states of the participant);
- *cognitions* (attention to selected aspects of the current situation);
- *planning* (exploring sequences of possibilities mentally); and
- *evaluations* (explicit or implicit comparisons of alternatives).

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

If one translates a metaphor literally under the silent condition but paraphrases it under the TA condition, then TA has undoubtedly changed the translation process or cognitive structure. Obviously, this is not what the findings by Krings (2001) and Jakobsen (2003) are about.

TA has also been used in reading and writing studies. As translation involves reading and writing, related studies in these two fields merit our attention. Smith et al. (2019) provided a methodological review of 76 original empirical studies published between 2000 and 2015 that used verbal reports (including TAPs) to investigate the reading processes of language learners; six of these studies discussed the reactivity of verbal reports. Leow and Morgan-Short (2004) and Bowles and Leow (2005) found that verbalization had a minimal effect on text comprehension and written production and did not compromise the validity of the study. Godfroid and Spino (2015) reported that neither eye-tracking nor TA affected text comprehension, although TA had a small, positive effect on vocabulary recognition. Bowles (2010) did a meta-analysis of 14 studies and concluded that TAPs “can reliably be used as a data collection tool” (p. 138); meanwhile, she pointed out that reactivity is multidimensional and, depending on many factors, can improve or hinder task performance (p. 137).

Compared with the issue of reactivity of TA, veridicality has been a minor concern for research across all fields, that is, whether TAPs reflect thoughts accurately (Bowles, 2018). In a study among 43 Chinese sophomores, Yang (2019) compared TAPs with retrospective verbal reports in an English-as-a-Foreign-Language (EFL) writing task and reported that TAPs were largely accurate, although there were various insignificant omissions in them. Thought processes mediating revision decisions were thought aloud for only about 8% of the revisions, therefore the monitoring processes were mostly under-represented in the TAPs. Based on the participants' reflections, the omissions appeared to be covert (as intermediary processes), transient (a second or two), nearly automatized or intuitive, making them hard to verbalize. According to Ericsson and Simon's theory (1993, p. 90), only information in focal attention can be verbalized, while such omissions are not in focal attention. In other words, TAPs do not provide an exhaustive view of the cognitive processes that occur during a task (Jääskeläinen, 2017). This attests to the usefulness of retrospective verbal reports in providing complementary information such as elaborations, rationalizations and justifications, although such information is not what Ericsson and Simon (1993, p. xvi) recommend deriving from retrospective verbal reports for accuracy concerns.

### **3. Cognitive effort and its measurement**

The term “cognitive effort” is used in many fields and has many synonyms, including mental load, mental workload, cognitive workload, cognitive load, cognitive effort, mental effort, task difficulty (Sun, 2019). Mental workload (sometimes referred to as cognitive workload) has been an important concept in human factors and industrial psychology. It is defined in various ways, for example, “portion of the operator's limited capacity actually required to perform a

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

particular task” (O’Donnell & Eggemeier, 1986, p. 42-2). The term “cognitive load” has been used in psychology since the 1960s, and is defined by Block et al. (2010) in a meta-analytic review as

“the amount of information-processing (especially attentional or working memory) demands during a specified time period; that is, the amount of mental effort demanded by a primary task” (p. 331).

The term “mental load” first appeared in the 1920s and has been used interchangeably with mental workload in the field of psychology. “Cognitive effort” has been defined as “the engaged proportion of limited-capacity central processing” (Tyler et al., 1979, p. 607). “Difficulty”, from a cognitive perspective, refers to the amount of cognitive effort required to solve a problem (see Sun, 2015). From these definitions, we may or may not see the distinctions among them, but we can see the connections.

According to Muñoz Martín (2012), mental load is a pivotal construct in Cognitive Translation Studies (CTS); research in this regard can help us to identify important characteristics of the translation process, unravel the complex relationships between attention, consciousness, problem-solving, automation and expertise, and may be beneficial for both current theoretical and empirical efforts in translation process research. Similarly, Lacruz (2017) believes that studying cognitive effort is “key to gaining insights into the translation process” (p. 387).

A central theme in the study of mental load, cognitive effort and translation difficulty is how to measure them. In their bibliography on mental workload assessment, Wierwille and Williges (1980) identified 28 specific techniques in four major categories: subjective opinion, spare mental capacity, primary task and physiological measures. In translation process research, the applicable techniques can be classified into three major categories: subjective, performance and physiological measures (e.g., measures of brain, eye, cardiac or muscle functions). Subjective measures have been used frequently and the most commonly used subjective measure is the rating scale. A frequently employed rating scale is the Task Load Index (NASA-TLX) developed by Hart and Staveland (1988), which has six workload-related subscales: mental demand, physical demand, temporal demand, effort, performance, and frustration level. It has been adopted and adapted in translation research (see Sun & Shreve, 2014). The performance measures include time on task, typing behaviour (such as pauses, deletions) recorded by key-logging programs, eye movement and gaze direction (e.g., fixation count, fixation duration) revealed by eye-tracking and translation quality (Halverson, 2017; Lacruz, 2017). An important factor to consider in a study related to cognitive effort is the characteristics of the ST/SL or TT/TL, which have been found to correlate to cognitive effort (Halverson, 2017; Liu et al., 2019; Sun & Shreve, 2014) and may have an impact on the effect of TA on cognitive effort (Sun, 2011).

## 4. Methods

### 4.1 Participants

Twenty graduate students in Translation or Interpreting from a university in Beijing participated in this experiment in early 2019. Three of them took part in an ST difficulty evaluation, three in a pilot test and 14 were the informants in the main experiment. They were native Chinese speakers from 21 to 24 years of age. All the participants – bar one – were female. They all signed an informed consent form and each received about USD25 in compensation.

### 4.2 Test passages

Five passages were used in this study. One of them was in Chinese, about 100 words long, used in the copying task for participants to get used to the same keyboard used in all the experimental tasks. The other four passages were in English, were about 150 words long, and were used as STs in the translation tasks. These four self-contained passages were excerpts taken from two recently published texts entitled, respectively, “The Uses of Nostalgia” retrieved from *The Economist* and *Worldwide Cancer Rates Rising* from voanews.com.

Based on the Flesch Reading Ease formula, the *Nostalgia* piece was more difficult, with the two excerpts (i.e., ST1 and ST2) scoring 46.4 and 45.4 (corresponding to an 11th grade level), respectively; the *Cancer* piece was easier, with the two excerpts (i.e., ST3 and ST4) scoring 72 and 65.9 (corresponding to a 7th grade level), respectively. The translation difficulty level of the two excerpts taken from the same text in both cases was further confirmed to be approximately the same in terms of task duration and subjective evaluation by three participants. Therefore, two excerpts taken from the same text were homogeneous in terms of word count, text type, topic and ST difficulty.

Task analysis was conducted before the experiment. The English texts did not require specialized knowledge in order to be comprehended. Since the participants were not allowed to access the internet during the tests, English explanations for a few difficult words and Chinese translations for English proper nouns were shown to the participants on paper for a few minutes just before the test on the computer began.

### 4.3 Instruments and experimental setup

A Tobii TX300 eye tracker and a high-performance PC were used in the experiment; installed on the PC were the key logger Translog-II for recording keyboard and mouse events, Tobii Studio, and Audacity for audio recording.

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

In an eye-tracking experiment, the type and model of the eye tracker, the physical distance between eye tracker and participant's eyes, the type of chair for the participant, the participant's ability to touch type, light intensity, among others, may have an impact on the quality of the data collected (Hvelplund, 2014; O'Brien, 2009). Therefore, pilot studies are indispensable.

The participants were informed beforehand not to wear contact lenses, false eyelashes or eyeliners, which would affect the validity of the eye-tracking data (O'Brien, 2009). The distance between eye tracker and participant was around 64 cm, as recommended by the *Tobii Studio Manual* (p. 36). A comfortable non-swivelling, medium-height chair was selected.

The pilot study showed that a participant's sitting posture had an influence on the collection of eye-tracking data. Participants who moved their body forward and backwards from time to time when performing the task produced low-quality data. During one pilot session, a participant moved her chair and the quality of the eye-tracking data immediately became poor. Therefore, before an experiment started, we did eye-tracking calibration and adjusted the chair position to ensure the distance between eye-tracker and participant was optimal, and asked the participant to keep their posture steady while translating. The participant's eye-movement habits may result in data-collection failure: if a participant tends to blink very fast or look up and to the left (or right), they will be a poor fit for experiments with a remote eye tracker. This was confirmed in our pilot study.

The project setup in Translog-II was also an important factor to consider. For Tobii Studio to record the whole screen of Translog-II, the Translog-II User's window was adjusted out of full-screen mode. The text in the Translog-II window was double-spaced so that the participant's gaze path and fixations would be easier to recognize.

#### **4.4 Test procedure**

Each participant completed two sessions: first, they translated two passages (ST1 and ST3) silently and, second, they translated the other two passages (ST2 and ST4) while TA (see Table 1). Behaviour at the keyboard and mouse was recorded with Translog-II, which does not track translation behaviour outside its interface (e.g., web searches), so the participants were not allowed to access the internet or refer to print materials.

In each session, the translation task was preceded by the previously noted copying task and the two texts were presented in random order. The sessions had no time constraints. For each participant, there was a two-day interval between the two sessions. The participants were told that the quality of their translations would be assessed and therefore they could review their translation and make revisions if necessary when they had finished the draft and before submitting it. This was supposed to push participants to try their best in the experiment. The

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

participants were not informed of the focus of this study, that is, the impact of TA on the measures of cognitive effort.

**Table 1: Texts and translation conditions**

ST	ST difficulty	Translation condition
1	difficult	silently
2	difficult	thinking aloud
3	easy	silently
4	easy	thinking aloud

After the participants had finished the translation of a passage, they were asked to complete an adapted NASA-TLX survey for evaluating the translation difficulty level of the passage, which comprised four items with English and Chinese descriptions: mental demand, effort, frustration and performance. Each category was rated on a 0–10 scale, with 0 being extremely low and 10 being extremely high (see Sun & Shreve, 2014).

Before the TAPs session, the participants received some training on TA: based on Ericsson and Simon’s (1993, p. 376) recommendations, they were asked to multiply two numbers in their head and say out loud everything that they would say to themselves silently; they were then requested to translate a few sentences from English into Chinese on paper while thinking aloud so that they would grow accustomed to the experiment. In addition, the participants were told to focus on the translation task, especially when facing a choice between TA and performing the translation task (Sun, 2011), and to avoid articulating what they assumed the observer wanted to hear (Jääskeläinen, 2017). After they had finished the translation task, each of them was interviewed about the difficulties they had encountered with their translation and the influence of TA on their translation process and performance. After all the tests were completed, the participants’ translations were assessed with regard to quality by two graders. Participants’ verbalizations and post-performance interviews were then transcribed and analysed. These verbal data were analysed together with key-logging and eye-tracking data.

#### **4.5 Data quality and analysis**

Two post-experiment measures were adopted to ensure data quality: Tobii Studio's indicators of the recording quality and Translog-II’s replay function. Tobii Studio presented several types of information after each recording, including the ID of the recording, the name of the participant, date, duration, Gaze Samples and Weighted Gaze Samples, the last two of which are indicators of the number of valid samples during a recording.

According to the Tobii Studio User’s manual (v. 3.4.8), the Gaze Samples show how many of the eye-tracking samples in a recording have usable gaze data and indicate how useful the



Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

recording will be for analysis. It is expressed as a percentage that is “calculated by dividing the number of eye-tracking samples with usable gaze data that were correctly identified, by the number of attempts” (p. 41); 100% means that one or two eyes were detected for the full recording and 50% indicates that one or two eyes were detected for half of the recording session. A low value means that the participant’s eyes could not be found because they were looking away from the screen or were closed during parts of the recording. As human beings cannot open their eyes uninterruptedly, Gaze Samples cannot be 100% in an authentic experiment. The Weighted Gaze Samples value is “weighted based on if both or just one eye was detected in a sample” (p. 41); 100% shows that two eyes were detected for the full recording whereas 50% indicates that one eye was detected throughout the recording or two eyes for a half of the recording.

Compared to Gaze Samples, Weighted Gaze Samples seem to be a more precise measure. Therefore, Weighted Gaze Samples were adopted and those recording sessions with a Weighted Gaze Samples percentage below 60% were replayed and examined in Translog-II. In the end, data from one of the 14 formal participants were discarded and only 13 participants’ data were analysed.

The Translog-II keylogging files were uploaded to an online TPR-DB management tool, which tokenized the STs and TTs, segmented them into sentences, and aligned them at a sentence level. After some adjustments, the sentence-aligned STs and TTs were manually aligned at a word level. The TPR-DB management tool then generated TPR-DB tables, which were downloaded to the local computer for statistical analysis. These TPR-DB tables (see Figure 1) contained more than 200 features that can be used to describe and model translation behaviour, such as timestamps, durations, pauses, deletions, insertions, part of speech tags, number of typed keystrokes per activity unit, number of fixations per activity unit and word translation entropy (see Carl et al., 2016).

Session	SL	TL	Part	Text	Task	Dur	TimeD	TimeR	FixS	TrtS	FixT	TrtT	Ins	De
P19_T07	en	zh	P19	1	T	925110	11266	698407	907	372916	961	423114	1251	96
P21_T07	en	zh	P21	1	T	1344047	3547	1284485	1354	596712	1226	470873	1274	87
P30_T07	en	zh	P30	1	T	1393610	2828	972110	907	416904	1162	480779	1246	111
P31_T07	en	zh	P31	1	T	911703	5531	622656	746	309868	913	464256	1546	181
P32_T07	en	zh	P32	1	T	933360	42719	854360	519	266626	672	428372	1321	93
P34_T07	en	zh	P34	1	T	1445797	6453	1338406	765	344004	1331	653741	1249	64
P35_T07	en	zh	P35	1	T	1721110	10422	1093141	1313	395708	2037	557523	992	70
P36_T07	en	zh	P36	1	T	859094	2782	777000	580	218584	1190	427495	1240	121
P39_T07	en	zh	P39	1	T	934016	3922	779063	653	334775	816	380219	1417	74
P41_T07	en	zh	P41	1	T	1358609	90890	888187	443	207196	504	181186	1333	111
P42_T07	en	zh	P42	1	T	662828	9203	438610	453	217990	759	326115	1072	29
P44_T07	en	zh	P44	1	T	1277015	128500	824593	1088	372066	1471	677640	1347	121
P45_T07	en	zh	P45	1	T	1431688	137516	1191094	1241	666325	1259	535022	1353	83
P19_T08	en	zh	P19	2	T	892359	14906	850265	866	351755	964	397257	1233	111
P21_T08	en	zh	P21	2	T	1502860	5250	940657	1718	727839	1169	470655	1074	76
P30_T08	en	zh	P30	2	T	1542266	3219	1119266	1442	665143	1023	394165	1229	88

Figure 1 Screenshot of part of a TPR-DB table

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

The data were analysed using R, a statistical software package (R Core Team, 2018), and descriptive statistics and linear mixed-effects modelling were performed on the data with such R packages as lmerTest (Kuznetsova et al., 2019). In this study, the fixed effects were the TA condition and the ST difficulty, each with two levels. The random effect was the participant factor. The dependent variables comprised:

- duration of tasks;
- duration of the orientation phase;
- duration of the draft phase;
- duration of the revision phase;
- fixation duration;
- fixation count; and
- translation difficulty level.

Data from eye-tracking and key-logging were analysed together with TAPs.

## 5. Results and discussion

### 5.1 Effects of TA on translation duration

**Table 2 Descriptive statistics on translation duration (in seconds)**

Conditions	Mean	Median	Skewness	Kurtosis
Silently	1,017.07	933.69	.40	-.90
TA	1,308.33	1,298.16	.69	.13
Easy ST	1,063.76	1,015.80	1.00	.10
Difficult ST	1,261.64	1,298.16	.56	.24
Total	1,162.70	1,122.69	.70	.40
<i>Note. N = 52.</i>				

Table 2 shows that the mean duration for translating a passage in the formal experiment was 1,162.7 s; under the silent condition, it was 1017.07 s, while under the TA condition it was 1,308.33 s. That is, the duration under the TA condition was about 30% longer than that under the silent condition. The mean duration of easier tasks (i.e., translating texts 3 or 4) was 1,063.76 s, whereas for more difficult tasks (i.e., translating texts 1 or 2), the mean was 1,261.64 s. That is, in this study, translating difficult STs took about 20% longer than translating easier STs.

Skewness and kurtosis, two measures of the symmetry in a distribution, were found to be 0.7 and 0.4, respectively, which indicated the distribution of data was moderately right-skewed. Two outliers with durations of 2,201.55 and 2,267.45 s were found with a box plot; they

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

exerted great influence on the variance duration of ST2 translations, and therefore were excluded from the inferential statistical analysis. A density plot then showed that the data of duration of 50 observations (i.e., 13 participants × 4 passages – 2 outliers) was normally distributed. Linear mixed-regression modelling was employed to estimate the influence of TA on translation.

**Table 3 Translation duration (fixed-effects model estimates)**

	Estimate	SE	95% CI		t	p
			LL	UL		
Intercept	865.07	76.04	709.48	1020.65	11.377	<.001
TA	341.59	69.64	200.63	481.11	4.905	<.002
Diff	304.01	67.72	167.73	440.29	4.489	<.003
TA:Diff	-210.08	97.75	-406.78	-13.37	-2.149	.038
<i>Note.</i> N = 50.						

Table 3 shows that the *t* value of the fixed effects of TA was 4.905 ( $p < 0.01$ ), indicating that TA had a significant influence on translation duration; the translation duration under the TA condition was estimated by a linear mixed-regression model to be 341.59 s longer than that under the silent condition without considering the random effect of the participants. This finding is consistent with Ericsson and Simon’s theory (1993, p. xxxii) that participants slow down only moderately due to the additional verbalization and also with Jakobsen’s (2003) finding that TA delayed translation by about 25%.

The *t* value of fixed effects of ST difficulty was found to be 4.489 ( $p < 0.01$ ), showing that ST difficulty had a significant influence on translation duration; the duration for translating a difficult ST was estimated to be 304.01 s longer than for translating an easier ST.

The interaction between the TA condition and ST difficulty condition was also found to be significant; the *t* value of fixed effects of the interaction was -2.149 ( $p < 0.01$ ). The duration difference in the translation of easier texts between the silent condition and the TA condition was estimated to be 341.59 s; for more difficult texts, the difference was only 130.93 s. From this, it can be inferred that when the level of ST difficulty increases, the difference in translation duration between the silent condition and the TA condition will decrease.

## 5.2 Effects of thinking aloud on different translation phases

In Translog data, the translation process was divided into orientation, draft and revision phases (Dragsted & Carl, 2013; Nitzke & Oster, 2016). The orientation phase is that during which a translator reads the ST before any insertion or deletion takes place. It lasts from the beginning of a translation session until the first keystroke. The draft phase starts from the end of the orientation phase until the last token of the ST is translated. The revision phase refers to the

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

period during which a translator deletes and/or inserts words in the TT after having produced a complete draft, and it starts with the first deletion or insertion. It should be mentioned that this division of phases is based on key demarcation points and it is not intended to filter out instances of revision taking place prior to a complete draft being in place.

### 5.2.1 On the duration of the orientation phase

The mean duration of the orientation phase ( $n = 50$ ) was 21.32 s, median 8.51 s and standard deviation (SD) 30.7 s. The minimum duration of the orientation phase was 2.53 s, whereas the maximum was 137.98 s. The skewness and kurtosis were found to be 2.31 s and 4.83 s, respectively, indicating the distribution of data was heavily right-skewed and there were outliers. A box plot (see Figure 2) analysis of the data showed ten outliers.

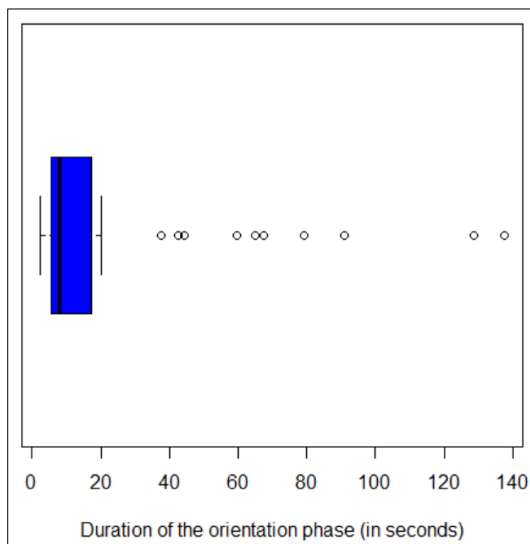


Figure 2 Outliers in the duration data

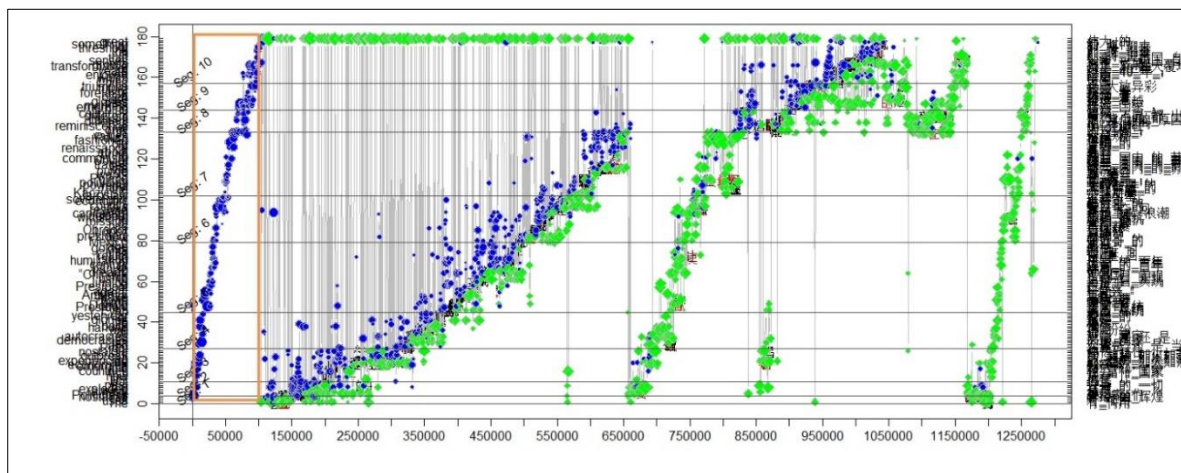


Figure 3 Translation progression graph of one participant translating a passage

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

In a Translation Progression Graph (see Figure 3), the horizontal axis shows the time (in milliseconds (ms)) when the translation of the ST was produced; the left vertical axis enumerates the emerging ST words whereas the right vertical axis gives the target words; blue dots represent fixations on the ST; green dots represent fixations on the TT; black characters represent insertions; red characters represent deletions (Carl et al., 2016). The orientation phase in Figure 3 is marked by an orange rectangle. As we can see, the participant read all the sentences of the ST before typing the TT.

Such great discrepancies among the observations pointed to participants' different translation styles. During the orientation phase, several types of behaviour could be distinguished via the eye-tracking data:

- (1) reading through the ST before translation;
- (2) skimming the ST quickly before translation;
- (3) reading the first couple of words or sentences before pressing the first key (Carl et al., 2011).

Since the distribution of data was heavily right-skewed, the data were transformed and normalized by taking the logarithm before statistical analysis. A linear mixed model was built ( $t = -0.804$ ,  $p = 0.426 > 0.05$ ) and it indicated that the fixed effect of TA was not significant. Then, ST difficulty was added into the model and the  $t$ -value was 0.431 ( $p = 0.669 > 0.05$ ); it showed that the fixed effects of ST difficulty were not significant either. The interaction between TA and ST difficulty was not significant ( $p = 0.693 > 0.05$ ). Therefore, the duration of orientation was not influenced significantly by TA, ST difficulty or the interaction of both. This makes sense, as the participants might simply be reading the ST out loud.

### 5.2.2 On the duration of the draft phase

The mean duration of the draft phase ( $n = 50$ ) was 901.89 s, median 860.59 s and SD 322.13 s. The data were transformed and normalized by taking the logarithm before statistical analysis. A linear mixed model comprising two fixed effects (TA, ST difficulty) and one random effect (participant) was applied. Both fixed effects ( $t = 3.825$  for TA condition,  $p = 0.0004 < 0.01$ ;  $t = 2.141$  for ST difficulty,  $p = 0.039 < 0.05$ ) were found to be significant on the duration of the draft phase. The interaction between TA and ST difficulty had no significant influence ( $p = 0.693 > 0.05$ ) on the duration of the draft phase. In other words, TA increased the duration of the draft phase and, compared to translating easier texts, translating difficult texts required a longer duration in this phase.

### 5.2.3 On the duration of the revision phase

The mean duration of the revision phase ( $n = 50$ ) was 239.49 s, median 223.11 s and SD 168.59 s. The minimum duration was 6.19 s, whereas the maximum was 627.97 s. The

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

distribution of data was slightly right-skewed. A linear mixed model comprising two fixed effects (TA, ST difficulty) and one random effect (participant) was applied. The results showed that the fixed effect of TA was not significant ( $t = 1.696$ ,  $p = 0.097 > 0.05$ ); the fixed effect of ST difficulty was significant ( $t = 2.426$ ,  $p = 0.020 < 0.05$ ); the interaction between the two predictor variables had no significant influence ( $p = 0.693 > 0.05$ ) on the duration of the revision phase. In other words, as the ST difficulty level increased, the duration of the revision phase also increased. The best model estimated that the duration of the revision phase for difficult texts would be 89.97 s longer than that for easier texts and the standard error was 39.38 s; the duration of the revision phase under the TA condition would be 65.30 s longer than under the silent condition, and the standard error was 39.38 s.

As to random effects, the residual variance – that is, the variability that was left unexplained by the fixed effects in the model – in this case (the residual variance was 22,861) – was much larger than the participant variance. This means that other factors were affecting the duration of the revision phase, such as a participant's translation style, time pressure, fatigue and even hunger.

### 5.3 Effects of thinking aloud on processing the STs and TTs

The exploration of the effects of TA on translation needs to take cognitive effort into account. In Translog data, cognitive effort can be measured by fixation count and fixation duration. Because, text wise, translation consists of comprehending the ST and producing the TT, we need to study the effects of TA on processing the STs and TTs according to fixation count and fixation duration. In the TPR-DB tables,

“FixS and FixT are the number of fixations on the source token(s) and on the target token(s), while TrtS and TrtT represent the total reading time, i.e. the sum of all fixation durations on the source and target text respectively” (Carl et al., 2016, p. 21).

We replayed the data collected with Translog-II and found that data from two participants contained little fixation and therefore excluded them (i.e., eight translations by the two participants) from our statistical analysis.

#### 5.3.1 Fixation count and duration on the ST

The mean of FixS ( $n = 44$ ) was 978, median 907 and SD 327.51. A linear mixed model comprising two fixed effects (TA, ST difficulty) and one random effect (participant) was applied. Both fixed effects ( $t = 5.172$  for TA condition,  $p = 7.28e-06 < 0.01$ ;  $t = 3.241$  for ST difficulty,  $p = 0.002 < 0.01$ ) were found to be significant on FixS. The interaction between TA and ST difficulty had no significant influence on FixS ( $p > 0.05$ ). In other words, TA had significant influence on the number of fixations on the ST and, based on the estimate of fixed effects, the count of fixations on the ST would increase by 345 under the TA condition as opposed to under the silent condition; ST difficulty also had a significant influence on the number of fixations on the ST and, compared

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

to translating easier texts, translating difficult texts would require 166 more fixations on the ST. Even when the two passages are at the same difficulty level, TA would increase the count of fixations on the ST.

The same finding applies to fixation duration on the ST. The mean of TrtS ( $n = 44$ ) was 442.51 s, and median 387.46 s. In the linear mixed model constructed, both fixed effects ( $t = 7.185$  for TA condition,  $p = 1.2e-08 < 0.01$ ;  $t = 4.237$  for ST difficulty,  $p = 0.0001 < 0.01$ ) were found to be significant on TrtS. The interaction between TA and ST difficulty was not significant ( $p > 0.05$ ).

### 5.3.2 Fixation count and duration on the TT

The mean of FixT ( $n = 44$ ) was 1,051, median 1,011 and SD 286.47. A linear mixed model comprising two fixed effects (TA, ST difficulty) and one random effect (participant) was applied. The fixed effect of TA ( $t = 0.069$ ,  $p = 0.94 > 0.05$ ) was not significant on FixT, whereas the fixed effect of ST difficulty ( $t = 3.129$ ,  $p = 0.003 < 0.01$ ) was significant on FixT. The interaction between TA and ST difficulty had no significant influence on FixT ( $p > 0.05$ ).

The mean of TrtT ( $n = 44$ ) was 486.13 s, median 451.05 s and SD 134.47 s. In the linear mixed model constructed, both fixed effects ( $t = 2.581$  for TA condition,  $p = 0.006 < 0.01$ ;  $t = 2.183$  for ST difficulty,  $p = 0.035 < 0.05$ ) were found to be significant on TrtT. The interaction between TA and ST difficulty had no significant influence on TrtT ( $p > 0.05$ ).

In other words, the participants spent more time looking at the TT under the TA condition than under the silent condition; when working on a difficult ST, the participants had more fixations on and spent more time looking at the TT compared with working on an easier ST. The findings that participants spent more time looking at both the ST and TT under the TA condition than under the silent condition are also in line with previous findings that TA would prolong the translation duration by about 30% compared to the silent condition.

### 5.4 Cross-check with the adapted NASA Task Load Index

The adapted NASA-TLX includes four subscales and represents the participant's subjective evaluation of the translation difficulty level of a passage. Descriptive statistical analysis of NASA-TLX scores showed that the trimmed mean (TM) of translations under the TA condition (TM = 5.34) was greater than that under the silent condition (TM = 4.76) and that the trimmed mean of translations of two more difficult STs (TM = 5.5) was greater than that of two easier STs (TM = 4.6).

A linear mixed model comprising two fixed effects (TA, ST difficulty) and one random effect (participant) was applied. Both fixed effects ( $t = 4.139$  for TA condition,  $p = 0.0001 < 0.01$ ;  $t = 5.551$  for ST difficulty,  $p = 2.17e-06 < 0.01$ ) were found to be significant on NASA-TLX scores. The interaction between TA and ST difficulty was significant ( $p = 0.029 < 0.05$ ).

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

In other words, TA had a significant influence on the perceived translation difficulty level. The NASA-TLX score was estimated by the best model to increase by 0.947 from 4.124 under the silent condition to 5.071 under the TA condition. In addition, the NASA-TLX score for translating difficult texts would be 1.270 higher than for translating easier texts. TA would reduce the effect of ST difficulty on NASA-TLX score (TA: ST difficulty =  $-0.732$ ); that is, when the task changes from translating an easy ST under the silent condition to translating a difficult ST under the TA condition, the NASA-TLX score would increase by 1.485 (i.e.,  $0.947 + 1.270 - 0.732$ ). In a word, the effects of TA on translation duration and translation difficulty level were consistent. Therefore, TA increased the cognitive effort in translation.

### **5.5 Cross-check with the post-experiment interview**

When a participant finished their two translation sessions, they were interviewed in Chinese and asked three questions. In answer to the first question about whether they felt comfortable with TA, almost all of them said that their thoughts were faster than their verbalizations and that TA interfered with their translation speed. This was consistent with our statistical finding that TA had a significant influence on translation duration.

In response to the second question about the influence of TA on their translation process, their responses varied. Some participants reported that TA helped them with checking whether the translation was natural and in making mindful decisions about which words should be used. Others responded that TA tended to dampen sparks of ideas and insights and increase their frustration. Yet others replied that TA interfered with their translating at the beginning, and then they got used to it. Two participants said TA had not influenced their translation process.

The third question was concerned with the impact of TA on translation quality, and here the participants had differing opinions, ranging from positive to negative to that TA had had no influence. Some believed that it had no influence on translation quality when they were working on a difficult ST. The results from our statistical analysis of translation quality scores showed TA had no influence on the quality of the translation of difficult STs ( $p = 0.957 > 0.05$ ), but lowered the quality of the translation of easier STs ( $p = 0.013 < 0.05$ ). This was, to some extent, consistent with Bowles and Leow's (2005) finding that verbalization did not affect text comprehension or written production significantly, although metalinguistic verbalization appeared to cause a significant decrease in text comprehension as opposed to non-metalinguistic verbalization.

## **6. Conclusion**

This study investigated the effects of TA on cognitive effort in translation according to time on task, the duration of different translation phases (i.e., orientation, draft, revision), cognitive



Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

effort involved in processing the STs and TTs, and translation difficulty as functions of ST difficulty level. The results were as follows:

- (1) TA has a significant influence on translation duration, and the duration under the TA condition is about 30% longer than that under the silent condition.
- (2) The effects of TA on cognitive effort vary as a function of ST difficulty: when the level of ST difficulty increases, the difference in translation duration between the silent condition and the TA condition will decrease.
- (3) The duration of the cognitive effort in the draft phase is increased due to TA, but the effects of TA on the duration of the orientation phase and the revision phase were not significant.
- (4) The cognitive effort expended on ST processing and TT processing according to fixation duration is increased due to TA.
- (5) TA significantly increases the perceived level of translation difficulty as measured by NASA-TLX.
- (6) The participants' post-experiment interviews indicate that the effects of TA on their translation process may differ.
- (7) TA had no influence on the quality of the translation of difficult texts, but slightly reduced the quality of the translation of easier texts.

This study did not take into account such factors as the directionality of translation and translation expertise (novices vs professionals), both of which probably have effects on the results (see Jakobsen, 2003); we leave these factors to future research. It would seem that, at this stage, sweeping generalizations such as “TA changes the course or structure of cognitive processes” will not help or convince people, for there are many observed and latent variables involved. We recommend investigating the specific effects of TA on the translation process, for example, whether TA makes participants “reluctant to make large-scale lexical changes, like omissions or additions” (Jääskeläinen, 2011, p. 20), so that researchers can avoid those “disturbing” factors through careful research design and as a result elicit valid data.

### **Acknowledgements**

We wish to thank the editors and two anonymous reviewers for their valuable comments and suggestions and the experiment participants for their time and effort.

### **Funding**

This work was supported by Beijing Social Science Foundation (grant number 16YYB017) and the Fundamental Research Funds for the Central Universities (grant number 2020JS001).

Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.

## References

- Angelone, E. (2010). Uncertainty, uncertainty management, and metacognitive problem solving in the translation task. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 17–40). John Benjamins. <https://doi.org/10.1075/ata.xv.03ang>
- Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, 134(3), 330–343. <https://doi.org/10.1016/j.actpsy.2010.03.006>
- Bowles, M. A. (2010). *The think-aloud controversy in second language research*. Routledge. <https://doi.org/10.4324/9780203856338>
- Bowles, M. A. (2018). Introspective verbal reports: Think-alouds and stimulated recall. In A. Phakiti, P. DeCosta, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 339–357). Palgrave Macmillan. [https://doi.org/10.1057/978-1-137-59900-1\\_16](https://doi.org/10.1057/978-1-137-59900-1_16)
- Bowles, M. A., & Leow, R. P. (2005). Reactivity and type of verbal report in sla research methodology: Expanding the scope of investigation. *Studies in Second Language Acquisition*, 27(3), 415–440. <https://doi.org/10.1017/S0272263105050187>
- Carl, M., Bangalore, S., & Schaeffer, M. (Eds.). (2016). *New directions in empirical translation process research*. Springer. <https://doi.org/10.1007/978-3-319-20358-4>
- Carl, M., Dragsted, B., & Jakobsen, A. L. (2011). A taxonomy of human translation styles. *Translation Journal*, 16(2). <https://translationjournal.net/journal/56taxonomy.htm>
- Carl, M., Schaeffer, M., & Bangalore, S. (2016). The CRITT translation process research database. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New directions in empirical translation process research* (pp. 13–54). Springer. [https://doi.org/10.1007/978-3-319-20358-4\\_2](https://doi.org/10.1007/978-3-319-20358-4_2)
- Dragsted, B., & Carl, M. (2013). Towards a classification of translation styles based on eye-tracking and key-logging data. *Journal of Writing Research*, 5(1), 133–158. <https://doi.org/10.17239/jowr-2013.05.01.6>
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215–251. <https://doi.org/10.1037/0033-295X.87.3.215>
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press. <https://doi.org/10.7551/mitpress/5657.001.0001>
- Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive?: A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137(2), 316–344. <https://doi.org/10.1037/a0021663>
- Godfroid, A., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning*, 65(4), 896–928. <https://doi.org/10.1111/lang.12136>
- Halverson, S. L. (2017). Multimethod approaches. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 195–201). Wiley Blackwell. <https://doi.org/10.1002/9781119241485.ch11>

- Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.
- Hansen, G. (2005). Experience and emotion in empirical translation research with think-aloud and retrospection. *Meta*, 50(2), 511–521. <https://doi.org/10.7202/010997ar>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hvelplund, K. T. (2014). Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data. *MonTI. Monografías De Traducción E Interpretación*, 2014(1), 201–223. <https://doi.org/10.6035/MonTI.2014.ne1.6>
- Jääskeläinen, R. (2010). Think-aloud protocol. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of translation studies* (Vol. 1, pp. 371–373). John Benjamins. <https://doi.org/10.1075/hts.1.thi1>
- Jääskeläinen, R. (2011). Back to basics: Designing a study to determine the validity and reliability of verbal report data on translation processes. In S. O'Brien (Ed.), *Cognitive explorations of translation* (pp. 15–29). Continuum.
- Jääskeläinen, R. (2017). Verbal reports. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 213–231). Wiley-Blackwell. <https://doi.org/10.1002/9781119241485.ch12>
- Jakobsen, A. L. (2003). Effects of think aloud on translation speed, revision, and segmentation. In F. Alves (Ed.), *Triangulating translation* (pp. 69–95). John Benjamins. <https://doi.org/10.1075/btl.45.08jak>
- Jakobsen, A. L. (2017). Translation process research. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 21–49). Wiley Blackwell. <https://doi.org/10.1002/9781119241485.ch2>
- Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Rand McNally.
- Krings, H. P. (2001). *Repairing texts: empirical investigations of machine translation post-editing processes* (G. Koby, G. Shreve, K. Mischerikow & S. Litzer, Trans.). Kent State University Press.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2019). Package lmerTest: Tests in linear mixed effects models. <https://cran.r-project.org/web/packages/lmerTest/lmerTest.pdf>
- Lacruz, I. (2017). Cognitive effort in translation, editing, and post-editing. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 386–401). Wiley-Blackwell. <https://doi.org/10.1002/9781119241485.ch21>
- Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, 26(1), 35–57. <https://doi.org/10.1017/S0272263104026129>
- Liu, Y., Zheng, B., & Zhou, H. (2019). Measuring the difficulty of text translation: The combination of text-focused and translator-oriented approaches. *Target*, 31(1), 125–149. <https://doi.org/10.1075/target.18036.zhe>
- Muñoz Martín, R. (2012). Just a matter of scope: Mental load in translation process research. *Translation Spaces*, 1(1), 169–188. <https://doi.org/10.1075/ts.1.08mun>
- Nitzke, J. (2019). *Problem solving activities in post-editing and translation from scratch*. Language Science Press. <https://doi.org/10.4324/9780429030376-5>

- Sun, S., Li, T., & Zhou, X. (2020). Effects of thinking aloud on cognitive effort in translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 132–151.
- Nitzke, J., & Oster, K. (2016). Comparing translation and post-editing: An annotation schema for activity units. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New directions in empirical translation process research* (pp. 293–308). Springer. [https://doi.org/10.1007/978-3-319-20358-4\\_14](https://doi.org/10.1007/978-3-319-20358-4_14)
- O'Donnell, R. D., & Block, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Vol. 2. Cognitive Processes and Performance* (pp. 42-1–42-49). Wiley.
- O'Brien, S. (2009). Eye tracking in translation-process research: Methodological challenges and solutions. In I. M. Mees, F. Alves, & S. Göpferich (Eds.), *Methodology, technology and innovation in translation process research* (pp. 251–266). Samfundslitteratur.
- R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org>
- Smith, P., Kim, D., Vorobel, O., & King, J. R. (2019). Verbal reports in the reading processes of language learners: A methodological review. *Review of Education*, 8(1), 37–114. <https://doi.org/10.1002/rev3.3170>
- Sun, S. (2011). Think-aloud-based translation process research: Some methodological considerations. *Meta*, 56(4), 928–951. <https://doi.org/10.7202/1011261ar>
- Sun, S. (2015). Measuring translation difficulty: Theoretical and methodological considerations. *Across Languages and Cultures*, 16(1), 29–54. <https://doi.org/10.1556/084.2015.16.1.2>
- Sun, S. (2019). Measuring difficulty in translation and post-editing: A review. In D. Li, V. L. C. Lei, & Y. He (Eds.), *Researching cognitive processes of translation* (pp. 139–168). Springer Singapore. [https://doi.org/10.1007/978-981-13-1984-6\\_7](https://doi.org/10.1007/978-981-13-1984-6_7)
- Sun, S., & Shreve, G. M. (2014). Measuring translation difficulty: An empirical study. *Target*, 26(1), 98–127. <https://doi.org/10.1075/target.26.1.04sun>
- Toury, G. (1991). Experimentation in translation studies: Achievements, prospects and some pitfalls. In S. Tirkkonen-Condit (Ed.), *Empirical research in translation and intercultural studies* (pp. 45–66). Narr.
- Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, 5(6), 607–617. <https://doi.org/10.1037/0278-7393.5.6.607>
- Vieira, L. N. (2017). Cognitive effort and different task foci in post-editing of machine translation: A think-aloud study. *Across Languages and Cultures*, 18(1), 79–105. <https://doi.org/10.1556/084.2017.18.1.4>
- Wierwille, W. W., & Williges, B. H. (1980). *An annotated bibliography on operator mental workload assessment (Naval Air Test Center Report No. SY-27R-80)*. Naval Air Test Center.
- Yang, C. (2019). The veridicality of think-aloud protocols and the complementary roles of retrospective verbal reports: A study from EFL writing. *The Asia-Pacific Education Researcher*, 28(6), 531–541. <https://doi.org/10.1007/s40299-019-00453-5>