Cognitive load and cognitive effort: Probing the psychological reality of a conceptual difference

Anne Catherine Gieshoff
Andrea Hunziker Heeb

Zurich University of Applied Sciences

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

Cognitive load and cognitive effort: Probing the psychological reality of a conceptual difference

The cognitive demands associated with performing a task involve at least two dimensions: 1) the load dimension that is related to the assumed task difficulty and 2) the effort dimension that reflects the resources invested in a task. This study considers whether this distinction is actually relevant to translators and interpreters when they report load and effort and, if so, how the assumed psychological reality of these two dimensions is related to task performance. In this study, professional translators and interpreters performed naturalistic tasks with comparable stimuli, working from English into German. After each task, they were asked to rate their experienced load and effort as part of the NASA Task Load Index. Their performance was measured by analysing process and product indicators that correspond in interpreting and translation. Results indicate that while self-reported load and effort are highly correlated, their relationships to process or product measures appear to be more complex.

Keywords: cognitive load; cognitive effort; translation professionals; interpreting professionals; task performance; NASA Task Load Index

1 Introduction

Translation and interpreting are considered cognitively demanding tasks, as they require cognitive processes such as monitoring, retrieval of possible translation equivalents and decision taking or sentence planning (for an overview see Angelone et al 2016; for a definition of 'cognitive process' see APA 2023). The intensity with which these cognitive processes are performed seems to be related to different factors that are inherent either to the task or to the task performer. Indeed, the investigation of such factors has sparked a large body of research in cognitive translation and interpreting studies (CTIS). In the last twenty years, around 200 publications exploring the effects of cognitive processing have appeared in CTIS (Gieshoff et al 2021). A wide range of terms and expressions have been used to refer to these effects (for a list see Hunziker Heeb et al 2021), but two seem to be particularly dominant: cognitive (or mental)[1] load (e.g., Chen 2017; Chmiel et al 2017; Muñoz 2012; Plevoets and Defrancq 2018; Seeber 2015) and cognitive (or mental) effort (e.g., Alves 2007; Gile 2009; Lacruz 2017). Although the terms are sometimes used interchangeably, we think it essential to differentiate between the two: "[C]ognitive load [is associated] with the complexity of the stimuli and task (i.e., source text, commission, situation, and so on), and cognitive effort with the actual response by the task performer" (Ehrensberger-Dow et al 2020: 221; see also Gieshoff 2021). This seems very similar to Gile and Lei's understanding, since they state that "'cognitive load' is used to denote the cognitive pressure that a process imposes by virtue of environmental and task-specific factors, while 'cognitive effort' refers to the effort actually expended by the Translator [sic] when performing the task" (2021: 275). In economic terms, load can be understood as corresponding to the price tag of a task whereas effort is the amount the person performing it is willing to pay. Notwithstanding the usefulness of this distinction at a conceptual level, the question is whether translators and interpreters actually make this distinction when they are asked to report their experienced level of load or effort. Moreover, it is unclear how these self-reported measures relate to measures of task performance. In order to investigate these questions, firstly, we compared self-reports on cognitive load and effort after a translation or interpreting task, and secondly, we analysed their relationship to measures of translation and interpreting performance. Before delving into the data analysis, we will first take a closer look at the notions of load and effort within and also beyond CTIS.

[1] According to the APA Dictionary of Psychology, the term *cognitive process* is often used interchangeably with *mental process* (APA 2023). Paas et al (2003: 64) define the concept of cognitive load as multidimensional with mental load, mental effort and performance as three measurable aspects. In this article, we have adopted the word choice of the authors when we refer to published frameworks or models, and otherwise employ the term *cognitive*.

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

1.1 Cognitive Load and Cognitive Effort

Research interest in the effects of cognitive activity dates back over a century. One of the earliest records is probably a study published in 1899 where the authors already used the term "mental effort" to explain physiological changes observed during mental activity (Angell and Thompson 1899). Since then, the relationship between cognitive activity and physiology on one hand and attention and task performance on the other has served as a starting point for theoretical frameworks focusing on cognitive effort. These frameworks essentially draw on the notion of cognitive control to explain the effortful nature of cognitive tasks (Kahneman 2013; Shenhav et al 2017). Cognitive control is required whenever a task is not automated and needs deliberate attention, for instance, inhibiting an automated response, updating information or holding information in memory. The expected task benefit (Shenhav et al 2017) or the performer's level of expertise (see Young et al 2015; Young and Stanton 2004) are assumed to impact the amount of cognitive control that a performer decides to exert during a task, which can influence task performance. With the introduction of technology and computers in the workplace in the 1980s, it became more and more important to identify factors that affect cognitive processing, not least to prevent accidents and fatal errors in safety-critical domains (Young et al 2015). The emphasis then shifted towards cognitive load, i.e., properties that make a task more or less prone to errors. Again, it has generally been assumed that cognitive processes compete for limited attentional resources, but here the focus lies more on task demands than internal factors. Examples for such models are Wickens' Multiple Resource Theory (Wickens 2008), Sweller's Cognitive Load Theory (2010; 2011), or the *time-based resource sharing model* developed by Barrouillet and colleagues (Barrouillet et al 2007; Barrouillet et al 2004). Despite inconsistent use of terminology and a lack of precise definitions (see also Hunziker Heeb et al 2021), this basic distinction can also be observed in CTIS: load models highlight task-related aspects, like the amount of interference inherent in a task (Seeber 2015; 2011), whereas effort models emphasise internal factors that determine the allocation of cognitive resources (Alves and Gonçalves 2013; Gile 2009; Pym 2015).

In translation studies, Krings' (2001) three different types of effort—temporal, technical and cognitive—have been adopted by many other researchers, although the relation between the effort types is not straightforward and seems to depend on the individual translator's expertise and working style (Lacruz 2017). For this reason and since the cognitive acts of translating—and interpreting—and their manifestation in observable activities during task performance are so closely interwoven, we refrain from distinguishing between different types of effort. Considering that CTIS revolves around load and effort, it is quite surprising that so few authors have investigated both effort and load (for exceptions, see for instance Sun 2015; Chen 2017).

This raises an intriguing question: Can we disentangle load and effort empirically? In the following, we will discuss the relation between load and effort in two methods that have been commonly used in CTIS: self-reports and performance measures, which we divide into process and product data for the purpose of this study.

1.2 The Relationship between Load and Effort in Self-Report Measures

Although—to the best of our knowledge—no empirical study in CTIS has investigated the relationship between self-reported load and effort so far, the question has already sparked some hypothetical reflections. The relationship is generally assumed to be positive albeit not necessarily linear in the sense that study participants presumably adapt their effort subconsciously to the perceived task demands (see for instance Chen 2017; Ehrensberger-Dow et al 2020). This assumption is also largely in line with Kahneman (1973: 15f). It is also possible that study participants do not differentiate between load and effort or implicitly assess their effort when asked about task demands or load. In both cases, the conceptual differences underpinning load and effort will be difficult to distinguish in self-reports since perceived effort would increase with perceived load (unless, of course, an individual simply gives up).

Against this backdrop, it is interesting to note that the NASA Task Load Index (NASA-TLX, Hart and Staveland 1988), one of the most widely used instruments to collect self-reports on

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

workload, includes both dimensions of load[2] and effort. The NASA-TLX was developed in multiple experiments to identify those factors that significantly contribute to self-reported overall workload while being relatively independent from each other. A recent meta-review found a positive though only weak to moderate correlation between load and effort ($\kappa = 0.57$, $p < 0.01$)[3], which suggests that both dimensions are perceived by respondents as similar but not identical (Hertzum 2021: 5). Differences between self-reported effort and load are rarely observed, but they have been reported, for instance, with the use of incentives in a working memory task (Jang et al 2020) and with different levels of alertness in a driving simulation (Galy et al 2018).

In CTIS, it has been suggested in particular that high task demands may place a translator or interpreter in a situation of overload where the perceived demands exceed the available resources (Gile 2009). In such a situation, translators or interpreters may—at least in theory—rate their load considerably higher than their effort. While the concept of overload has been intensely studied in CTIS, the idea of underload (i.e., a situation where task demands are supposed to be low but performance is still suboptimal; Young et al 2015) has received much less attention, probably because interpreting and translation have traditionally been regarded as complex tasks. Still, it is conceivable that in such a situation participants will retrospectively feel that their effort was not well adapted to their load, either because they did not expend a sufficient amount of effort to perform at an optimal level or because they had to exert more effort than expected to reach a satisfactory level of performance. Cases where effort and load ratings differ may thus be particularly informative when it comes to the relationship between effort and load.

1.3 The Relationship between Load or Effort and Performance Measures

Probably the most commonly cited hypothesis on effort and performance in CTIS is Gile's "tightrope hypothesis" (Gile 2009; 2017), which assumes that interpreters work close to their saturation levels and predicts a decrease in performance (i.e., errors or infelicities) with any further increase in task demands. In the field of ergonomics and psychology, a different framework has been adopted: performance follows in general the shape of an inversed U with optimum performance in performers' individual 'comfort zone' whenever effort is close to load. A decrease in performance is observed in situations of underload and overload, that is in situations where effort is not well adapted to load (Young et al 2015). Specific strategies and skill can help to cope with task demands (Gile 2017; Young et al 2015). Empirical evidence, however, is less clear. In his meta-review, Hertzum (2021) found that the error rate increased with higher load or effort ratings, but the association was only weak (load: $\kappa = 0.19$, effort: $\kappa = 0.23$). This is in accordance with findings in CTIS. Sun and Shreve (2014) reported only weak correlations between translation quality as assessed by raters ($r = -0.12$) or time on task ($r = 0.29$) and a shortened version of the NASA-TLX[4]. It must be borne in mind, however, that in both cases the NASA-TLX score did not always distinguish between the task manipulations in the experiment. Hertzum (2021) notes that effort and load ratings did not differ between the experimental conditions in 174 of 245 tests. Likewise, Sun and Shreve (2014) found only a weak correlation (tau = -0.14) between readability scores and the shortened NASA-TLX version. Hence, it is possible that the conditions were not sufficiently distinct from each other to cause changes in NASA-TLX scores.

Another explanation for the weak correlation between load and performance may be that effort compensates at least to some degree for high load. If this is true, product quality can remain stable with increasing load until a certain limit is reached when the performer can no longer exert the required effort. This relationship is probably more pronounced in translation since time constraints are less important. A translator can at least in theory achieve a high-quality

[2] The item is called "mental demand" in the questionnaire. It was derived from the subjectively experienced task difficulty and represents the load dimension. The other items are effort, physical demand, temporal demand, performance satisfaction and frustration.

[3] In 590 tests, the NASA-TLX did not show any significant differences between the experimental manipulations. According to the author, the reason is probably that the conditions in these tests were too similar.

[4] Temporal and physical demand were excluded since they were "not applicable" according to the authors (Sun and Shreve 2014, 104).

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

translation of a very technical or complex text by taking the time to research terms and definitions and consult parallel texts. In this case, effort and load may only affect process measures but not product measures. This is different for interpreters: they may have the opportunity to prepare for a topic when very technical or dense talks are expected (see Díaz-Galaz et al 2015), but they do not have any possibility to do in-depth research during an assignment. This means that translation product and process measures might react differently to effort and load, this difference might be absent in interpreting since the process is reflected in the product. With this in mind, we decided to use process as well as product measures to measure performance for the purpose of this study. We will use the term *performance measures* to refer to both product and process measures.

Our aim was to choose indicators that are as similar as possible between translation and interpreting to allow for a comparison between the effects of load and effort on performance in both tasks. Product measures were defined in terms of product quality. For the interpretations, quality was operationalised as interpreting accuracy, that is the degree to which the meaning in the output is consistent with the meaning in the source talk. This approach was chosen since 'sense consistency' is a central quality criterion for interpreters and their listeners alike (Zwischenberger 2010: 130; Kurz 2002). As regards translation, product quality was operationalised as a two-faceted construct consisting of accuracy (correspondence of meaning in the source and target text) and fluency (the idiomaticity of the target text), adapted from Koehn and Monz (2006). The reason to include an additional aspect to accuracy was an expected ceiling effect: We assumed that translators may reach a high level of accuracy since time constraints are less strict than in interpreting. As it has been suggested that long pauses in particular may be related to difficulties in source text processing in interpreting (Setton 1999: 245–48) and translation alike (Muñoz & Cardona 2019), they were chosen as suitable process measures that the activities have in common. In interpreting, silent pauses are an important predictor of perceived fluency (Yu and van Heuven 2017). In translation studies, it has recently been suggested that long pauses be reframed not as problem indicators per se but as instances of strategic management of cognitive resources to help re-establish focus and attention (Angelone and Marín 2022: 68). Long pauses, whether they are taken deliberately or not, occur of course throughout the translation process. However, we decided to concentrate on the translation drafting phase in order to increase comparability with the flow of language production in the interpreting task.[5]

2. Study

The data we present in this section was collected as part of the project *Cognitive load in interpreting and translation*.[6] The project primarily investigates whether processing non-standard English language input increases the cognitive demands of interpreters, translators and other multilinguals compared to processing a version of the same input that has been edited to conform to the conventions of standard English (for details on the editing see Ehrensberger et al 2020). For the purpose of this study, we focused on the edited versions only. We were interested in three research questions (R1-R3) with four related hypotheses (H1-H4), as outlined below.

- R1: What is the relationship between self-reported load and effort in translation and interpreting?
  H1: Based on the literature, we expect that both self-reported load and effort in translation and interpreting will correlate well and will be within the individual comfort zone of each participant. Instances where ratings differ may be indicative of underload or overload.
- R2: How are self-reported load and effort related to product measures?
  H2: We hypothesise that product measures will follow an inverse U-shape, indicating that product quality is suboptimal whenever effort is not well adapted to load.

[5] We acknowledge that this is a simplified approach to the translation process, which ignores inherent process activities such as online research and end revision that may take a substantial amount of process time (see for example Hvelplund 2017).

[6] For more information about the CLINT project see http://www.zhaw.ch/linguistics/iued/clint

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

- R3: How are self-reported load and effort related to process measures?
  H3: We assume that for translation, process measures are more strongly related to effort than to load. For interpreting, however, we do not expect any difference in the relationship between effort or load and process measures.
  H4: Furthermore, we assume that for translation, process measures are more closely related to self-reported effort than product measures are.

2.1 Participants

A total of 28 professional interpreters (henceforth called interpreters) and 24 professional translators (henceforth called translators) were recruited for the study. The interpreters' professional experience ranged from 2 to 38 years (MD_{years} = 20.5, *SD* = 12.38); the translators had 4 to 42 years of experience (MD_{years} = 13, *SD* = 9.82). All had German as their native language and English as one of their working languages. Each of them signed an informed consent form and filled in a questionnaire about their professional and linguistic background before participating in the study.

2.2 Stimuli

The edited versions of two authentic conference talks (henceforth called talks) and the corresponding abstracts written by the speaker of the talk and submitted to the conference organisers (henceforth called abstracts) were used as stimuli (henceforth called source texts) for the study. Table 1 summarises the main properties of the source texts. The topic of the first set of talk and abstract was mobility. It can be characterised as rather generic, with a low proportion of words that are not among the 5000 most frequent words of American English (Davies 2008). The second set about (economic) demand forecasting was highly technical and contained many expressions that are not among the 5000 most frequent words of American English (Davies 2008). In order to enhance comparability, both talks were respoken by a professional speaker of general North American English and recorded on video.

Table 1: Stimuli properties

Talks				
Stimulus	Number of words	Duration (min:sec)	Delivery speed (syllables/min)	Uncommon words (%)
General talk	1508	12:05	204.2	4.0
Technical talk	1427	12:07	214.0	11.4
Abstracts				
	Number of words	Number of words/sentence (mean)	Function words (%)	Uncommon words (%)
General abstract	181	5.4	9.2	7.0
Technical abstract	179	5.5	8.3	17.7

2.3 Procedure

In a usability laboratory equipped with an office workstation and a desktop computer, participants were tested individually. Each participant processed both the general and the technical source text; one of them in its original non-Standard English version and the other one in its edited version. The texts were counter-balanced for order. After each, the participants completed the NASA-TLX. They were guided through the procedure with instructions in German, their native language, which appeared on the computer screen. In both groups, participants started with a warm-up task to familiarise themselves with the equipment. For the interpreters, the warm-up task consisted of interpreting a 5-minute speech; for the translators, it consisted of a two-minute typing and online research task.

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

After a short break, the main task started. First, a brief with information on the speaker or the author, respectively, and the conference was displayed on the screen. As the interpreters—unlike the translators—did not have the possibility to do research on the internet, they received a short paper glossary with potentially problematic source text terms (7 for the general talk and 14 for the technical task) as well as pencil, paper and a marker pen. They were given the time to consult the glossary and prepare themselves for the interpreting task. Participants started the interpreting or translation task themselves by pressing a key. The translators were given a translation brief and the source text. For the translation of the technical abstract, they also received the translation of two key expressions. They were given 30 minutes to perform the task. Their translation processes were recorded with the keylogging software Inputlog (Leijten and Van Waes 2013), and their target texts were saved as Word documents. Interpreters' renditions were recorded and saved as audio files. After the interpreting or translation task, participants reported their workload by filling in the NASA-TLX questionnaire (Hart and Staveland 1988). The NASA-TLX includes six items that are rated on a scale from 0 to 20; the endpoints are marked with "very low" and "very high", respectively. For the purpose of the present study, we focus on the items “mental demand”, queried by the question "How mentally demanding was the task? and “effort”, queried by the question "How hard did you have to work to accomplish your level of performance?".

2.4 Data Processing and Analysis

To analyse the relationship between self-reported load and effort, we conducted a two-way interclass correlation over the NASA-TLX *mental demand* and *effort* ratings, with the *mental demand* rating corresponding to cognitive load. The 'raters' were treated as fixed effects since they corresponded to our study participants. To assess the similarity between both ratings, we also computed the distance between the effort and load rating (ELD) by subtracting the second from the first value. Positive distance values indicate that participants rated their effort higher than their load, whereas negative values suggest that load was rated higher than effort.

To test the relationship of self-reported load and effort on performance measures, we conducted multiple linear regressions on the performance measures with self-reported load, self-reported effort and the distance between effort and load as predictor variables. Each predictor was tested as a linear and a quadratic term in order to test the hypothesis that performance is optimal when effort is adapted to load. Additionally, an interaction with text was tested to account for the possibility that any effects were different for the different texts. A final model with text as the predictor variable was conducted to ensure that the dependent variables were sensitive to differences in texts. P-values were corrected for multiple testing according to Bonferroni (significance level: 0.004). Each model was visually inspected for influential data points (Cook's distance above 0.14 for interpreting and 0.16 for translation)[7] and, if influential data points were observed, run again without the data points in question.

Performance measures included interpreting accuracy or translation quality, respectively, and the number of particularly long (silent) pauses during the interpreting process or the translation drafting process, respectively. Interpreting accuracy was assessed according to the method described in Gieshoff and Albl-Mikasa (2022) and the total score per participant was computed. As two different talks were used as stimuli, the accuracy scores were recalculated as a ratio of the maximum possible score for each text to make the scores comparable for both talks. As regards the assessment of translation quality, in each abstract seven segments were chosen that seemed particularly salient with regards to potential translation problems. Subsequently, two raters scored accuracy of content and idiomaticity of formulations for each segment in each translation on a scale from 1 to 5. Inter-rater reliability on both quality criteria was only moderate (accuracy: ICC = 0.66, $p < 0.01$; fluency: ICC = 0.6, $p < 0.01$). Therefore, we decided to add up the single scores of each segment in each target text and to use the mean between the two raters as a dependent variable for all further analyses.

Depending on the type of task, pauses were obtained either from the keylogging file or from the audio recording of the rendition. For the translation task, we only considered pauses during the drafting phase, that is the time span from typing the first character of the target text to the last

[7] Cook's distance was calculated according to the following formula: cook's distance = $4/n$

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

character of the draft. In order to see whether long internet searches or the reading of parallel texts found online might lead to a distortion of the results, we analysed the data separately, i.e., with and without the pauses that occurred while the internet browser was logged as the active screen window. As regards the interpreting task, we used Praat (Boersma and Weenink 2013) for the automatic detection of silent pauses. Pauses before 1.5 seconds after talk onset and after the end of the talk were deleted as these can be assumed to be related to ear-voice span rather than cognitive processes. The minimum pause threshold was set to 500 ms since an analysis of 8 random one-minute samples with two raters suggested that the automatic detection worked less reliably on lower thresholds (manual corrections necessary at 100 ms: *M* = 17.2; 250 ms: *M* = 2.43; 500 ms: *M* = 1.06). As we were particularly interested in pauses that might reflect cognitive processing, we counted in each rendition or translation the number of pauses that qualified as outliers according to the following formula: 3rd quartile + 1.5 interquartile range. The formulas and respective thresholds are displayed in Table 2. The median and third quartile are given for orientation. All statistical analyses were conducted with the basic functions available in R (R Core Team 2020) and tidyverse (Wickham et al 2019); model effects were computed with the effects package (Fox 2003) and plotted with ggplot (Wickham 2016).

Table 2: Calculation and threshold values for pauses in translation and interpreting

	Description	Translation	Interpreting
Threshold	75th percentile +1.5* IQR	1724 ms[8]	2349 ms
Third quartile	75% of all observations are below	848 ms	1407 ms
Median	50% of all observations are below	424 ms	890 ms

3. Results

In the following, we present the results for the relationships between self-reported load and effort and process measures as well as product measures.

3.1 Relationship between Self-Reported Load and Effort

The descriptive statistics of the load and effort ratings are displayed in Table 3. According to a Mann-Whitney-U test, translators and interpreters alike tended to perceive the technical source text as inducing more effort (translators: median (*MD)* = 12; interpreters: *MD* = 18) and load (translators: *MD* = 13; interpreters: *MD* = 19) than the general one (effort: translators: *MD* = 9; interpreters: *MD* = 14; load: translators: *MD* = 11.5; interpreters: *MD* = 13). The difference in effort and load as revealed with a Mann-Whitney-U-test was significant for the interpreter group (effort: $U = 20.5$, $p < 0.001$; load: $U = 0$, $p < 0.001$) but not for the translator group (load: $W = 39.5$, $p = 0.063$; effort: $U = 61$, $p = 0.54$). Effort and load ratings were strongly correlated with an interclass correlation coefficient of 0.79 and a confidence interval ranging from 0.667 to 0.877 ($F(51, 51.4) = 8.62$, $p < 0.001$). The distance between effort and load ratings showed that interpreters tended to rate their effort higher than their load (*MD* = 1.0, *SD* = 2.91), whereas translators reported higher load than effort (*MD* = -0.5, *SD* = 3.15). However, for 73.1% of the participants both ratings were no more than two units away from each other (on a 20-point scale) so can be considered quite close.

An examination of the ratings that were more than two units away from each other revealed that discrepant values are more frequent for the general source text than the technical text. Four out of 28 interpreters and two out of 24 translators found the general text considerably more effortful than their load rating would have suggested, whereas one interpreter and two translators rated their load higher than their effort. One interpreter and four translators assessed the load related to the technical text higher than their effort. No participant rated their effort higher than their load when working with the technical text.

[8] The threshold was close to the value for long pauses (1752 ms) computed according to the formula given by Muñoz and Cardona (2019).

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

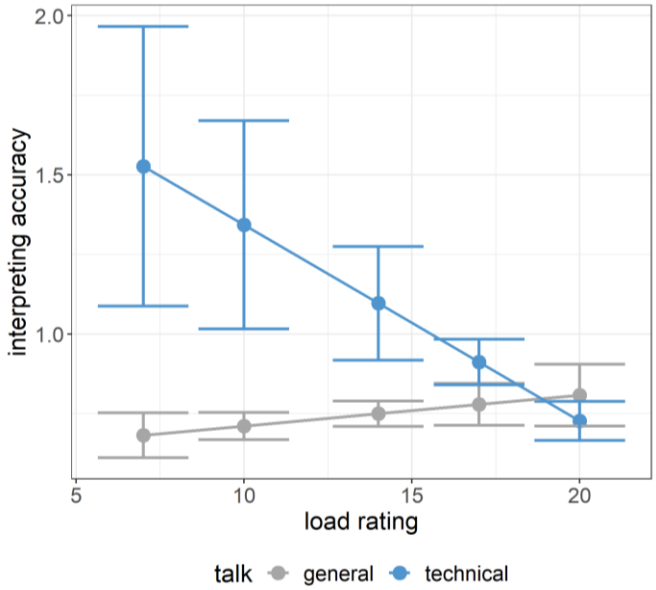
Table 3: Descriptive statistics for the NASA-TLX score, load score and effort score

		Mean	Median	SD	Skewness	Kurtosis
All	NASA score	58.04	53.50	24.47	0.28	2.23
	Load	13.40	13.50	4.76	-0.61	2.53
	Effort	13.17	13.17	4.97	-0.48	2.19
Translators	NASA score	42.00	41.50	16.58	0.43	3.14
	Load	11.33	12.50	4.34	-0.75	2.33
	Effort	10.08	11.00	4.44	-0.11	1.80
Interpreters	NASA score	71.78	73.50	21.72	-0.13	2.21
	Load	15.17	16.50	4.44	-0.86	2.75
	Effort	15.82	17.00	3.76	-0.99	3.46

3.2 Relationship between Load, Effort and Product Measures in Interpreting

The interpreting accuracy scores ranged from 0.64 to 0.91 indicating an overall acceptable accuracy (M = 0.77, SD = 0.084 skewness: -0.125, kurtosis: 1.765). Interestingly, the scores were on average higher for the technical talk (M=0.80, SD=0.081) than for the general talk (M = 0.74, SD = 0.077). Most models did not reach significance after correcting for multiple testing according to Bonferroni (significance level: 0.004; for the complete model statistics, see Table 4 in the appendix) and did not change after removal of influential data points. Two models were significant after removal of influential data points and after Bonferroni-correction for multiple testing (for the model statistics, see Table 5 in the appendix). The first model (three data points removed) indicated a significant main effect of talk (β = 1.343, SE = 0.344, t = 3.899, p < 0.001) and a significant interaction of talk and load rating (β = -0.071, SE = 0.019, t = -3.766, p = 0.001). The interaction suggests that interpreting accuracy decreases with higher load, but only for the technical talk (see Figure 2). With 47.9% variance explained, the model corresponds to a moderate effect size.

Figure 1: Effects of load rating on interpreting accuracy in both talks.[9]



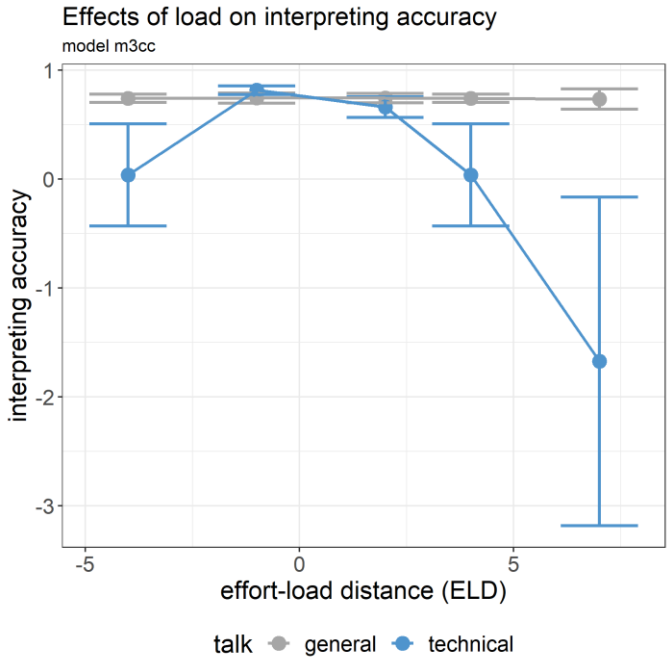
The second model (one data point removed) indicates a main effect of talk (β = 0.132, SE = 0.034, t = 3.820, p < 0.001), with an overall higher interpreting accuracy for the technical talk, and an interaction between the quadratic term of effort-load-distance and talk (β = -0.05,

[9] The error bars correspond to the standard error of the estimate in this figure and the following figures.

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

SE = 0.015, *t* = -3.596, *p* = 0.002). The negative estimate of the interaction term suggests an inverse U-shape for the technical talk, but not for the general one (see Figure 2). The model explains 35% of the variance, which can be described as a moderate effect.

Figure 2: Effects of effort-load distance (ELD) and talk on interpreting accuracy. The error bars correspond to the standard error of the estimate.

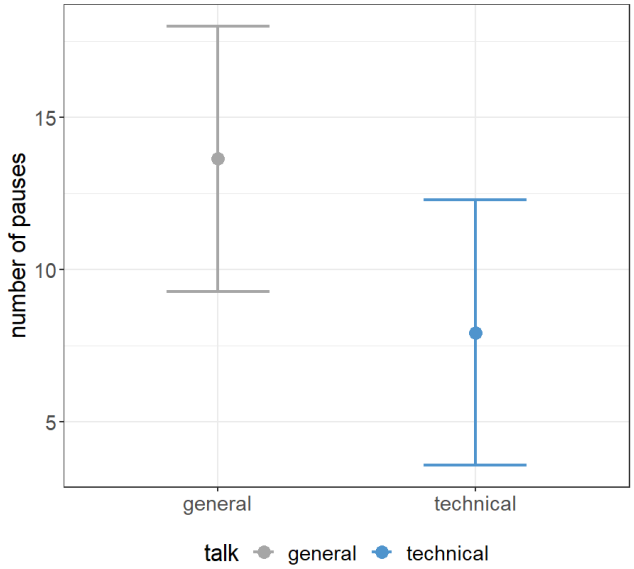


3.3 Relationship between Load, Effort and Process Measures in Interpreting

Between zero and 38 long silent pauses were observed in each interpreting rendition (*MD*=9.500, SD=8.316, skewness=1.804, kurtosis=6.346). The number of pauses was overall higher for the general talk (*M*=13.643, *MD*=12.000, *SD*=10.573) than for the technical talk (*M*=7.929, *MD*=8.000, *SD*=3.772). None of the models conducted on the number of long silent pauses reached significance (for the complete model statistics, see Table 6 in the appendix), although one model trended towards significance (*β*=-5.714, *SE*=3.000, *t* = --1.905, *p*=0.068). The negative estimate suggests that the number of long silent pauses was lower for the technical talk (see Figure 2). With 6.7% of the variance explained, the effect size of the model is very weak. None of the models changed significantly after removal of influential data points.

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

Figure 3: Effects of talk on the number of long pauses in interpreting



3.4 Relationship between Load, Effort and Product Measures in Translation

With a median of 61.5 points of a possible maximum of 70 points (*SD* = 2.19, skewness = -0.438, kurtosis = 2.867), translation quality was acceptable overall. It was slightly higher for the general abstract (*MD* = 61.25) than for the technical one (*MD* = 60.5), but this difference was not statistically significant (*U* = 85.5, *p* = 0.45). None of the tested predictors reached significance (see Table 7 in the appendix for the model statistics). The models' statistics did not change significantly after removal of influential data points.

3.5 Relationship between load, effort and process measures in translation

The number of long pauses during the drafting process ranged from 23 to 136 (*M* = 81.17, SD = 27.87, *MD* = 80, skewness = 0.089, kurtosis = 2.553) and was higher overall during translation of the general abstract (*MD* = 86.5) than the technical one (*MD* = 72.0), but the difference was not significant (*U* = 87, *p* = 0.402). Similar results were found when including pauses due to the translators' internet searches in the data (general: *MD* = 62.25; technical: *MD* = 60.75, *U* = 85.5, *p* = 0.45). No significant effect was observed on the number of long pauses for any of the predictors (see Table 8 and Table 9 in the appendix for the model statistics), even when running the same models with the internet searches included. The models' statistics did not change significantly after removal of influential data points.

4. Disentangling Effort and Load

In this study, we have explored the relationships between self-perceived effort and load on one hand and performance on the other, both in translation and interpreting. With regard to the differentiation of load and effort, we observed a strong correlation between self-reported load and effort, which supports our hypothesis H1 and is in line with Hertzum's (2021). In 73.1% of the cases, ratings of effort and load were less than two units from each other. As expected, the proximity between self-reported effort and load combined with the fact that 56.7% of all ratings were below the upper quarter of the scale (i.e., below 15 on a scale from 1 to 20) suggests that the interpreting or translation task was still within the comfort zone of most of the participants. In approximately one-quarter of the cases, however, both ratings differed quite substantially. Interestingly, translators tended to rate their load higher than their effort, whereas interpreters leaned more towards effort than load. The reason might be linked to time pressure: in interpreting, the source text is a continuous auditory input and requires a constant and immediate response. This may be perceived as effort by the interpreters whereas the translators had more

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

freedom to decide on their own rhythm to work on the task. Therefore, they may experience their cognitive demands as load rather than as effort.

With respect to our hypothesis H2, a relationship between self-reported cognitive demands and product measures was only observed in interpreting. Two different effects were observed: First, the product measure – operationalised as interpreting accuracy – decreased with higher load ratings. Second, performance was optimal when effort was best adapted to load; that is, when the effort ratings were close to the load ratings. Both observations may be seen very tentatively as support for the hypothesized U-shaped relationship between performance, effort and load (see H2; Young et al 2015). Yet the fact that this relationship was only observed for the technical talk casts some doubt on whether the interpretation of the results is really that simple. It is possible that other factors that were not examined in this study influenced product measures. Interpreters may have, for instance, adopted different approaches to optimise their rendition with regard to the perceived text function, which could weight quality criteria differently. Interpreters may have judged precision to be more important in the technical talk than the general one, with consequences for the accuracy scores. Furthermore, the number of cases where effort and load differed substantially was rather low, which increased the uncertainty for predicting these cases (i.e., as expressed in the rather large confidence intervals).

As regards the translation task, we were not able to observe any effects of self-reported load or effort related to translation quality in our data set. This is similar to Sun and Shreve (2014), who only found a very weak effect of NASA-TXL scores on translation quality (R^2=0.015). In our case, it is probably due to a ceiling effect in translation quality: translators had an average of 90% of the maximum score with little variance (SD = 2.19), which means that roughly two-thirds of the translators reached a score between 88% and 92%. At the same time, no differences in translation quality between the two abstracts were found, probably because the participants, all professionals, had enough time to do the translations in a satisfactory manner. Against this backdrop, we would suggest a simpler interpretation of the data in the sense that the task probably fitted the participants' competence level. Taken together, we only found partial support for hypothesis H2.

As regards the relationship between self-reported cognitive demands and process measures (hypothesis H3), we did not observe any statistically significant effects. The difference in technicality of the two talks was the only predictor that seemed to influence the number of long silent pauses in interpreting, even though this effect was not significant. Interestingly, the model suggests a lower number of long pauses in the technical talk than in the general one. The reason for this counter-intuitive result may again be found in interpreters' approach to both talks. The many technical terms in the technical talk may have required a shorter ear-voice span for an accurate and complete rendition, whereas the general talk may have left more room for interpretation and time for formulation.

Despite the differences in technicality of the two abstracts, we did not find any effects of self-reported effort or load on the number of long pauses in translation either with or without those associated with internet searches. This means that our hypotheses H3 and H4 are not supported. Nevertheless, a similar observation was reported by Hertzum (2021) who found in his meta-review that effort and load ratings did not distinguish between different levels of task difficulty in 174 out of 245 cases.

Another explanation might be related to the way pauses have been operationalised and the inherent differences between translation and interpreting as cognitive activities. In translation, pauses correspond to interruptions in the keyboard or mouse activity. This means that any activity that involves typing, scrolling or clicking it is not recorded as a pause. If a translator types in a word in an online dictionary, it does not lead to pauses in the keylogging activity. Unfilled pauses in interpreting, in contrast, are those moments when the interpreter stops talking. If an interpreter searches for a word in an online dictionary, the result will likely be in a pause in the rendition. Apart from a very short glossary, the participants in this study did not have the opportunity to consult the internet or related documents so unfilled pauses in the renditions can be more easily linked to difficulties in source text comprehension or target text production.

Another aspect is the persistence of the text input. Translation can of course involve a considerable amount of time pressure, but the entire source text is available to the translator at all times. This means that the translator can work and process the source text at their own rhythm. A translator can stop typing to read the source text or the target text without the risk of losing

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

information. In interpreting, by contrast, the constant speech input forces the interpreter to keep pace with the speaker. An interpreter who waits for new information in order to process larger information units risks overloading memory and, as a result, omitting important information. This implies that interpreters will generally try to avoid long pauses, which is not necessarily the case for translators. Long pauses might therefore not be a reliable indicator of cognitive activity in translation. At the same time, overly short pause durations may be problematic indicators in interpreting because they may simply mark word or sentence boundaries.

To summarize, our study participants did not seem to distinguish effort from load, or at least they had a very similar appreciation of both. In the same vein, we did not find clear-cut effects of self-reported effort or load related to performance. Against this background, it seems that the difference between effort and load was of little relevance to participants' experience of the cognitive demands of translation and interpreting, at least in our study. An approach to tackle this issue in future studies may consist of using a broader variety of source texts ranging from very easy to very difficult, ideally validated beforehand by professional translators' and interpreters' evaluations of the source text difficulty.

Acknowledgments

We would like to thank the anonymous reviewers and Maureen Ehrensberger-Dow for their constructive and helpful feedback. The project was funded by the Swiss National Science Foundation, grant number 173694.

References

Alves, Fabio. 2007. 'Cognitive Effort and Contextual Effect in Translation: A Relevance-Theoretic Approach'. *Journal of Translation Studies* 10 (1): 57–76.

Alves, Fabio, and José Luiz Gonçalves. 2013. 'Investigating the Conceptual-Procedural Distinction in the Translation Process'. *Target* 25 (1): 107–24. https://doi.org/10.1075/target.25.1.09alv

Angell, James Rowland, and Helen Bradford Thompson. 1899. 'A Study of the Relations between Certain Organic Processes and Consciousness.' *Psychological Review* 6 (1): 32–69. https://doi.org/10.1037/h0072367

Angelone, Erik, Maureen Ehrensberger-Dow, and Gary Massey. 2016. 'Cognitive Processes'. In *Researching Translation and Interpreting*, edited by Claudia V. Angelelli and Brian James Baer, 43–57. London: Routledge. http://www.tandfebooks.com/isbn/9781315707280

Angelone, Erik, and Álvaro Marín García. 2022. 'Reconceptualizing Breaks in Translation: Breaking Down or Breaking Through? *Translation & Interpreting* 14(2), Art. 2.

APA American Psychological Association. 2023. 'Cognitive Process'. *APA Dictionary of Psychology*. https://dictionary.apa.org/cognitive-process. Accessed 20 January 2023.

Barrouillet, Pierre, Sophie Bernardin, and Valérie Camos. 2004. 'Time Constraints and Resource Sharing in Adults' Working Memory Spans.' *Journal of Experimental Psychology: General* 133 (1): 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Barrouillet, Pierre, Sophie Bernardin, Sophie Portrat, Evie Vergauwe, and Valérie Camos. 2007. 'Time and Cognitive Load in Working Memory.' *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33 (3): 570–85. https://doi.org/10.1037/0278-7393.33.3.570

Boersma, Paul, and David Weenink. 2013. *Praat. Doing Phonetics by Computer* (version 5.3.51). http://www.praat.org.

Chen, Sijia. 2017. 'The Construct of Cognitive Load in Interpreting and Its Measurement'. *Perspectives* 25 (4): 640–57. https://doi.org/10.1080/0907676X.2016.1278026

Chmiel, Agnieszka, Agnieszka Szarkowska, Danijel Koržinek, Agnieszka Lijewska, Łukasz Dutka, Łukasz Brocki, and Krzysztof Marasek. 2017. 'Ear–Voice Span and Pauses in Intra- and Interlingual Respeaking: An Exploratory Study into Temporal Aspects of the Respeaking Process'. *Applied Psycholinguistics* 38 (5): 1201–27. https://doi.org/10.1017/S0142716417000108

Davies, Mark. 2008. 'Word Frequency Data'. *The Corpus of Contemporary American English* (COCA). https://www.english-corpora.org/coca/. Accessed 1 March 2022.

Díaz-Galaz, Stephanie, Presentación Padilla, and M. Teresa Bajo. 2015. 'The Role of Advance Preparation in Simultaneous Interpreting: A Comparison of Professional Interpreters and Interpreting Students'. *Interpreting. International Journal of Research and Practice in Interpreting* 17 (1): 1–25. https://doi.org/10.1075/intp.17.1.01dia

Ehrensberger-Dow, Maureen, Michaela Albl-Mikasa, Katrin Andermatt, Andrea Hunziker Heeb, and Caroline Lehr. 2020. 'Cognitive Load in Processing ELF: Translators, Interpreters, and Other Multilinguals'. *Journal of English as a Lingua Franca* 9 (2): 217–38. https://doi.org/10.1515/jelf-2020-2039

Fox, John. 2003. 'Effect Displays in R for Generalised Linear Models'. *Journal of Statistical Software* 8 (15): 27. https://doi.org/10.18637/jss.v008.i15

Galy, Edith, Julie Paxion, and Catherine Berthelon. 2018. 'Measuring Mental Workload with the NASA-TLX Needs to Examine Each Dimension Rather than Relying on the Global Score: An Example with Driving'. *Ergonomics* 61 (4): 517–27. https://doi.org/10.1080/00140139.2017.1369583

Gieshoff, Anne Catherine. 2021. 'Does It Help to See the Speaker's Lip Movements? An Investigation of Cognitive Load and Mental Effort in Simultaneous Interpreting'. *Translation, Cognition & Behavior* 4 (1): 1–25. https://doi.org/10.1075/tcb.00049.gie

Gieshoff, Anne Catherine, and Michaela Albl-Mikasa. 2022. 'Interpreting Accuracy Revisited: A Refined Approach to Interpreting Performance Analysis'. *Perspectives*. https://doi.org/10.1080/0907676X.2022.2088296

Gieshoff, Anne Catherine, Caroline Lehr, and Andrea Hunziker Heeb. 2021. 'Stress, Cognitive, Emotional and Ergonomic Demands in Interpreting and Translation: A Review of

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

Physiological Studies'. *Cognitive Linguistic Studies* 8 (2): 404–39. https://doi.org/10.1075/cogls.00084.gie
Gile, Daniel. 2009. *Basic Concepts and Models for Interpreter and Translator Training: Revised Edition*. 2nd ed. Vol. 8. Benjamins Translation Library. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/btl.8
Gile, Daniel. 2017. 'Testing the Effort Models' Tightrope Hypothesis in Simultaneous Interpreting - A Contribution'. *HERMES - Journal of Language and Communication in Business* 12 (23): 153. https://doi.org/10.7146/hjlcb.v12i23.25553
Gile, Daniel, and Victoria Lei. 2021. 'Translation, Effort and Cognition'. In *The Routledge Handbook of Translation and Cognition*, edited by Fabio Alves and Arnt Lykke Jakobsen, 1st ed. Routledge Handbooks in Translation and Interpreting Studies. New York: Taylor and Francis.
Hart, Sandra G., and Lowell E. Staveland. 1988. 'Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research'. In *Human Mental Workload*, edited by Peter Hancock and Meshkati, 52: 139–83. Advances in Psychology. North Holland: Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9
Hertzum, Morten. 2021. 'Associations among Workload Dimensions, Performance, and Situational Characteristics: A Meta-Analytic Review of the Task Load Index'. *Behaviour & Information Technology*, 1–13. https://doi.org/10.1080/0144929X.2021.2000642
Hunziker Heeb, Andrea, Caroline Lehr, and Maureen Ehrensberger-Dow. 2021. 'Situated Translators: Cognitive Load and the Role of Emotions'. In *Advances in Cognitive Translation Studies*, edited by Ricardo Muñoz Martín, Sanjun Sun, and Defeng Li, 47–65. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-2070-6_3
Jang, Hyesue, Ziyong Lin, and Cindy Lustig. 2020. 'Losing Money and Motivation: Effects of Loss Incentives on Motivation and Metacognition in Younger and Older Adults'. *Frontiers in Psychology* 11 (July): 1489. https://doi.org/10.3389/fpsyg.2020.01489
Kahneman, Daniel. 1973. *Attention and Effort*. Prentice-Hall Series in Experimental Psychology. Englewood Cliffs, N.J: Prentice-Hall.
Kahneman, Daniel. 2013. *Thinking, Fast and Slow*. 1st pbk. ed. New York: Farrar, Straus and Giroux.
Koehn, Philipp, and Christof Monz. 2006. 'Manual and Automatic Evaluation of Machine Translation between European Languages'. In *Proceedings of the Workshop on Statistical Machine Translation - StatMT '06*, 102. New York City, New York: Association for Computational Linguistics. https://doi.org/10.3115/1654650.1654666
Krings, Hans Peter. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Trans. Geoffrey S. Koby, Gregory M. Shreve, Katja Mischerikow, and Sarah Litzer. Kent, OH: Kent State University Press.
Kurz, Ingrid. 2002. 'Conference Interpreting: Quality in the Ears of the User'. *Meta* 46 (2): 394–409. https://doi.org/10.7202/003364ar
Lacruz, Isabel. 2017. 'Cognitive Effort in Translation, Editing, and Post-editing'. In *The Handbook of Translation and Cognition*, edited by John W. Schwieter and Aline Ferreira, 1st ed., 386–401. Wiley. https://doi.org/10.1002/9781119241485.ch21
Leijten, Mariëlle, and Luuk Van Waes. 2013. 'Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes'. *Written Communication* 30 (3): 358–92. https://doi.org/10.1177/0741088313491692
Muñoz Martín, Ricardo. 2012. 'Just a Matter of Scope. Mental Load in Translation Process Research'. *Translation Spaces* 1 (1): 169–88. https://doi.org/10.1075/ts.1.08mun
Muñoz Martín, Ricardo, and José Ma Cardona Guerra. 2019. 'Translating in Fits and Starts: Pause Thresholds and Roles in the Research of Translation Processes'. *Perspectives*, 27 (4): 525–551. https://doi.org/10.1080/0907676X.2018.1531897
Paas, Fred, Juhani E. Tuovinen, Huib Tabbers, and Pascal W. M. Van Gerven. 2003. 'Cognitive Load Measurement as a Means to Advance Cognitive Load Theory'. *Educational Psychologist* 38 (1): 63–71. https://doi.org/10.1207/S15326985EP3801_8
Plevoets, Koen, and Bart Defrancq. 2018. 'The Cognitive Load of Interpreters in the European Parliament: A Corpus-Based Study of Predictors for the Disfluency *Uh(m)*'. *Interpreting. International Journal of Research and Practice in Interpreting* 20 (1): 1–32. https://doi.org/10.1075/intp.00001.ple

Pym, Anthony. 2015. ‘Translating as Risk Management’. *Journal of Pragmatics* 85 (August): 67–80. https://doi.org/10.1016/j.pragma.2015.06.010
R Core Team. 2020. *R: A Language and Environment for Statistical Computing* (version 3.6.3). Vienna, Austria. https://www.R-project.org.
Seeber, Kilian G. 2011. ‘Cognitive Load in Simultaneous Interpreting’. *Interpreting* 13 (2): 176–204. https://doi.org/10.1075/intp.13.2.02see
Seeber, Kilian G. 2015. ‘Cognitive Load in Simultaneous Interpreting: Measures and Methods’. In *Benjamins Current Topics*, edited by Maureen Ehrensberger-Dow, Susanne Göpferich, and Sharon O’Brien, 72:18–33. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/bct.72.03see
Setton, Robin. 1999. *Simultaneous Interpretation: A Cognitive-Pragmatic Analysis*. Benjamins Translation Library 28. Amsterdam: Benjamins.
Shenhav, Amitai, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L. Griffiths, Jonathan D. Cohen, and Matthew M. Botvinick. 2017. ‘Toward a Rational and Mechanistic Account of Mental Effort’. *Annual Review of Neuroscience* 40 (1): 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526
Sun, Sanjun. 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, Sanjun, and Gregory M. Shreve. 2014. ‘Measuring Translation Difficulty: An Empirical Study’. *Target. International Journal of Translation Studies* 26 (1): 98–127. https://doi.org/10.1075/target.26.1.04sun.
Sweller, J. 2011. ‘Cognitive Load Theory’. *Psychology of Learning and Motivation - Advances in Research and Theory* 55: 37–76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8
Sweller, John. 2010. ‘Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load’. *Educational Psychology Review* 22 (2): 123–38. https://doi.org/10.1007/s10648-010-9128-5
Wickens, Christopher D. 2008. ‘Multiple Resources and Mental Workload’. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50 (3): 449–55. https://doi.org/10.1518/001872008X288394
Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Use R! Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4
Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al 2019. ‘Welcome to the Tidyverse’. *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686
Young, Mark S., Karel A. Brookhuis, Christopher D. Wickens, and Peter A. Hancock. 2015. ‘State of Science: Mental Workload in Ergonomics’. *Ergonomics* 58 (1): 1–17. https://doi.org/10.1080/00140139.2014.956151
Young, Mark S., and Neville Anthony Stanton. 2004. ‘Mental Workload’. In *Handbook of Human Factors and Ergonomics Methods*, edited by Neville Anthony Stanton, Alan Hedge, Karel Brookhuis, Eduardo Salas, and Hal W. Hendrick, 0 ed. CRC Press. https://doi.org/10.1201/9780203489925
Yu, Wenting, and Vincent J. van Heuven. 2017. ‘Predicting Judged Fluency of Consecutive Interpreting from Acoustic Measures: Potential for Automatic Assessment and Pedagogic Implications’. *Interpreting* 19 (1): 47–68. https://doi.org/10.1075/intp.19.1.03yu.
Zwischenberger, Cornelia. 2010. ‘Quality Criteria in Simultaneous Interpreting: An International vs. a National View’. *The Interpreters’ Newsletter* 15: 127–42

Appendix

Table 4: Model statistics (F-statistics, degrees of freedom, p-value and adjusted R value) for interpreting accuracy before removal of influential data points

predictor	model	F	DF	p-value	R_{adj}
effort rating	effort	0.226	1,26	0.638	-0.029
	effort*talk	1.859	3,24	0.164	0.087
	effort^2	0.371	1,26	0.548	-0.024
	effort^2*talk	1.838	3,24	0.167	0.085
load rating	load	2.300	1,26	0.142	0.045
	load*talk	2.373	3,24	0.095	0.132
	load^2	2.716	1,26	0.111	0.059
	load^2*talk	2.423	3,24	0.096	0.136
effort-load distance (ELD)	ELD	2.878	1,26	0.102	0.065
	ELD*talk	1.874	3,24	0.161	0.089
	ELD^2	0.596	1,26	0.447	-0.015
	ELD^2*talk	1.395	3,24	0.26	0.042
talk	talk	4.429	1,26	0.045	0.113

Table 5: Model statistics of significant models (F-statistics, degrees of freedom, p-value and adjusted R value) for interpreting accuracy after removal of influential data points

predictor	model	data points removed[10]	F	DF	p-value	R_{adj}
load rating	load*talk	3 (0.603, 0.202, 0.172)	8.365	3,21	< 0.001	0.479
effort-load distance (ELD)	ELD^2*talk	1 (224)	5.742	3,23	0.004	0.353

Table 6: Model statistics (F-statistics, degrees of freedom, p-value and adjusted R value) for the number of particularly long silent pauses in interpreting

predictor	model	F	DF	p-value	R_{adj}
effort rating	effort	1.195	1,26	0.284	0.007
	effort*talk	1.213	3,24	0.327	0.023
	effort^2	1.921	1,26	0.178	0.032
	effort^2*talk	1.294	3,24	0.299	0.031
load rating	load	1.636	1,26	0.212	0.023
	load*talk	1.271	3,24	0.307	0.029
	load^2	2.307	1,26	0.141	0.046
	load^2*talk	1.271	3,24	0.307	0.029
effort-load distance (ELD)	ELD	0.267	1,26	0.609	-0.028
	ELD*talk	1.193	3,24	0.334	0.021
	ELD^2	0.619	1,26	0.438	-0.011
	ELD^2*talk	1.842	3,24	0.166	0.085
talk	talk	3.628	1,26	0.068	0.088

Table 7: Model statistics (F-statistics, degrees of freedom, p-value and adjusted R value) for translation quality

independent variable	model	F	DF	p-value	R_{adj}
effort rating	effort	1.036	1,22	0.319	0.002
	effort*abstract	0.354	3,20	0.787	-0.092

[10] The numbers in brackets indicate the cook's distance of the data points that were removed.

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie

	effort^2	0.702	1,22	0.411	-0.013
	effort^2*abstract	0.267	3,20	0.848	-0.106
load rating	load	0.340	1,22	0.567	-0.029
	load*abstract	0.132	3,20	0.940	-0.128
	load^2	0.422	1,22	0.519	-0.025
	load^2*abstract	0.150	3,20	0.928	-0.125
effort-load distance (ELD)	ELD	0.373	1,22	0.548	-0.028
	ELD*abstract	0.296	3,20	0.828	-0.101
	ELD^2	0.509	1,22	0.483	-0.022
	ELD^2*abstract	0.381	3,20	0.767	-0.088
abstract	abstract	0.209	1,22	0.653	-0.034

Table 8: Model statistics (F-statistics, degrees of freedom, p-value and adjusted R value) for the number of particularly long pauses in translation

predictor	model	F	DF	p-value	R_{adj}
effort rating	effort	0.766	1,22	0.391	-0.010
	effort*abstract	0.787	3,20	0.516	-0.029
	effort^2	1.148	1,22	0.296	0.006
	effort^2*abstract	0.829	3,20	0.493	-0.023
load rating	load	0.006	1,22	0.940	-0.045
	load*abstract	0.297	3,20	0.827	-0.101
	load^2	0.015	1,22	0.905	-0.045
	load^2*abstract	0.391	3,20	0.761	-0.086
effort-load distance (ELD)	ELD	1.298	1,22	0.267	0.013
	ELD*abstract	1.521	3,20	0.240	0.064
	ELD^2	1.669	1,22	0.210	0.028
	ELD^2*abstract	1.039	3,20	0.397	0.005
abstract	abstract	0.874	1,22	0.360	-0.006

Table 9: Model statistics (F-statistics, degrees of freedom, p-value and adjusted R value) for the number of particularly long pauses in translation, including internet searches

predictor	model	F	DF	p-value	R_{adj}
effort rating	effort	0.48	1,22	0.496	-0.023
	effort*abstract	0.282	3,20	0.838	-0.103
	effort^2	0.850	1,22	0.367	-0.007
	effort^2*abstract	0.390	3,20	0.762	-0.086
load rating	load	1.229	1,22	0.28	0.010
	load*abstract	0.519	3,20	0.674	-0.067
	load^2	0.283	1,22	0.600	-0.032
	load^2*abstract	0.276	3,20	0.842	-0.104
effort-load distance (ELD)	ELD	0.547	1,22	0.467	-0.020
	ELD*abstract	0.464	3,20	0.711	-0.075
	ELD^2	3.297	1,22	0.083	0.091
	ELD^2*abstract	1.697	3,20	0.200	0.083
abstract	abstract	2.356	1,22	0.139	0.056

This is an Accepted Manuscript of an article published by John Benjamins in *Translation, Cognition and Behaviour* on 14 March 2023. The Version of Record is available online: https://doi.org/10.1075/tcb.00073.gie