Christian Olalla-Soler, Nicoletta Spinolo** & Ricardo Muñoz Martín****

Under Pressure? A Study of Heart Rate and Heart-Rate Variability Using the SmarTerp CAI tool

Abstract

The results of a quasi-experimental, intra-subject study are reported on the effects of the use of SmarTerp on physiological stress levels of twelve second-year students of the MA in Interpreting at the University of Bologna during a simultaneous interpreting task. The study, part of a broader project, explores the rendition of terminological units, proper names, and numbers and its correlation with stress levels, to provide insights into SmarTerp's practical usefulness in the field. Physiological stress levels were measured through heart rate and heart-rate variability indicators with Empatica E4 wristbands. Participants took part in three data-collection sessions over a month. In sessions 1 and 3 the participants interpreted two speeches, one with SmarTerp and another one without it. Descriptive findings hinted at a potential stress-alleviating effect of interpreting with SmarTerp, especially when interpreting into a second language. However, all inferential statistical results consistently revealed non-significant outcomes. Furthermore, stress levels did not decrease significantly over time when using SmarTerp. While the non-significant reduction in stress may cast doubt on the tool's efficacy, the complexity and multiple variables influencing stress in interpreting tasks should be factored in. SmarTerp may serve its primary purpose in aiding accurate rendition of terminological units, proper names, and numbers.

Keywords

interpreting; CAI tools; SmarTerp; quasi-experiment; physiological stress; heart rate; heart-rate variability

Introduction

The development of simultaneous interpreting technologies and the Internet has led to a technological breakthrough, which Fantinuoli (2018a) called "technological turn" in interpreting. This leap forward, predicted to impact all aspects of the profession, is based on advancements in remote interpreting, machine interpreting, and **computer-assisted interpreting** (CAI)—an oral translation form where interpreters use computer applications to support and enhance certain interpreting tasks and improve quality and productivity (Fantinuoli 2018b, p. 155). CAI tools support knowledge acquisition, terminology management, and entry retrieval, and they seem to reduce cognitive effort (Stoll 2009) and facilitate accurate rendering of terminological units (Díaz-Galaz 2015; Gacek 2015; Biagini 2016; Wang & Wang 2019) and numbers (Desmet et al. 2018; Defrancq & Fantinuoli 2021) when at task.

CAI applications can be categorized into first- and second-generation tools (Fantinuoli 2018b). First-generation tools emphasize terminology management, whereas second-generation tools further incorporate information retrieval from corpora, terminology extraction, and organization of materials. Recent advancements in simultaneous interpreting technologies have resulted in cloud-based **Remote Simultaneous Interpreting** (RSI) platforms or **Simultaneous Interpreting Delivery Platforms** (SIDPs), where interpreters operate with virtual consoles in virtual environments (Braun 2019; Saeed et al. 2022).

SmarTerp is a hybrid remote interpreting system and AI-powered computer-assisted interpreting tool combining RSI system functionality with CAI tool capabilities. It uses an artificial intelligence engine to provide live, automatic speech recognition of pre-selected problematic units: terms, proper

Christian Olalla Soler
Facultat de Traducció i Interpretació
Universitat Autònoma de Barcelona
christian.olalla@uab.cat

***Nicoletta Spinolo*
Dipartimento di Interpretacione
e Traduzione
Università di Bologna, Forlì
nicoletta.spinolo@unibo.it

****Ricardo Muñoz Martín*
Dipartimento di Interpretacione
e Traduzione
Università di Bologna, Forlì
ricardo.munoz@unibo.it

names, and numbers. SmarTerp basically works as follows: (1) the interpreter feeds problematic units into a glossary before the session; (2) when interpreting, live, automatic speech recognition of those units prompts their target language renditions on the screen. That is, SmarTerp identifies pre-selected units in the speech flow, locates them in the glossary, and prompts the interpreter with their translations from the glossary, potentially enhancing accuracy and coherence while reducing cognitive effort and stress (Prandi 2018, 2023; Frittella 2022).

The problematic units are elements on which boothmate support is usually expected (Gile 2009: 202). This becomes particularly relevant in RSI where, although not absent, remote boothmate support may become less effective (Chmiel & Spinolo 2022). On the other hand, however, the literature on visual input and RSI suggests that managing multiple input streams can be challenging and stressful for interpreters (Ziegler & Gigliobianco 2018; Saeed et al. 2023). Hence, interpreters' stress might increase due to introducing an additional stream of information and the need for interpreters to familiarize themselves with the tool, which imposes further processing demands.

Using tools like SmarTerp may have additional, unintended consequences for the interpreting process. Interpreters may continue to monitor the solutions being provided even when they are not or no longer needed, or when they decide to omit or generalize information. In such cases, prompts may become distractors. Other factors, such as a slow or unstable Internet connection, may lead to increased return lag, necessitating catch-up with the source speech after being prompted with the rendition. That is, interpreters may need to adjust their ear-voice span to the tool's or the connection's pace.

Using AI-powered CAI tools in the booth may have additional effects, due to the potentially more complex visual input to process if the interface is poorly designed or hinders human-computer interaction (Saeed et al. 2022). Despite these challenges, the stress induced by using such tools might decrease over time, as users become more familiar with them and learn to manage the additional information stream effectively.

The impact of tools like SmarTerp on the interpreting process may vary among interpreters, based on expertise levels; experienced interpreters might adapt more quickly, while trainees might experience increased cognitive demand (Wang & Wang 2019). This paper reports the results of a quasi-experimental, intra-subject study on the effects of using SmarTerp on physiological stress in a sample of second-year MA interpreting students at the University of Bologna. We aimed to answer two research questions:

(I) Does interpreting with SmarTerp reduce stress for MA interpreting students?
(II) Does exposure to SmarTerp decrease stress for MA Interpreting students over time?

While we acknowledge the inherent potential for heightened stress attributed to managing supplementary input streams, these tools may, in fact, mitigate stress over time. This alleviation may be due not only to the provision of ready-made solutions to potentially problematic units, but also to users adapting to the interface and the supplementary decision-making processes inherent to using the tool. As users become familiar with it, they might develop refined strategies and more efficient decision-making processes, which can in turn contribute to reducing stress levels. This assumption is rooted in our study, which involved students as participants and was developed within a didactic setup. We piloted SmarTerp at the University of Bologna to assess its usefulness and impact within a learning environment. Interpreting students are in the process of acquiring and refining their skills, so they offer a unique perspective and set of needs that are pivotal in evaluating the tool's efficacy and applicability in educational settings.

We posit that, if tools such as SmarTerp were to increase stress, including them in interpreting courses might warrant consideration. However, this decision should not be made hastily. The potential trade-off between increased stress and the possible enhancement of interpreting accuracy and overall performance should be factored in.

Besides, we cannot ignore the rapid advancement and growing prevalence of AI-powered CAI tools in the industry. Thus, including these technologies may be beneficial to prepare new cohorts of students to effectively navigate this evolving landscape. Early exposure to CAI tools should aid them in becoming more comfortable with such tools, thereby leveraging their advantages while simultaneously reducing the potential stress associated with their use. Over time, users' adaptation to these tools may be used as a valuable indicator of the tool's learnability, a key aspect of its overall usability (Nielsen 2010; Frittella 2023). To sum up, while adopting such tools might come with certain initial challenges, their long-term benefits could well outweigh the initial increase in stress levels, especially as users become more adept at managing the additional information stream and integrating it into their interpreting process.

To explore our research questions, we collected heart rate and heart rate variability data using Empatica E4 wristbands during interpreting tasks in two conditions, one with SmarTerp and another one without it, in three data-collection sessions over a period of a month. Section 2 presents the methods, with specific outlines for ethical issues (§2.1), data-collection procedure (§2.2), the sample (§2.3), the data-collection tool (§2.4), the indicators (§2.5), the stimuli (§2.6), and statistical procedures (§2.7). Section 3 presents and discusses the results and section 4 summarizes our main conclusions.

2. Materials and methods

2.1. Ethical aspects

The design of this study was approved by the Bioethics Committee of the University of Bologna. The informed consent form, signed by all participants, explained our goals and included information on the kind of data that would be collected (interpreters' performances and Empatica E4 data on cardiac activity). It also explained that only anonymized, aggregate data would be published. Participants were made aware that they could withdraw from the study at any moment and their data could be deleted with no consequences for them. No financial compensation was offered for their participation.

2.2. Procedure

The study was conducted in October 2021 in an interpreting lab of the Department of Interpreting and Translation of the University of Bologna, Forlì Campus.

Twelve second-year students of the MA in Interpreting participated in three data-collection sessions (one per week for all participants), in which they had to simultaneously interpret two speeches, one without SmarTerp and the other one with the tool, in that order. We were unable to randomize tasks and speech order due to practical constraints linked to SmarTerp running in its beta development phase at the time of data collection.

In both conditions, with and without SmarTerp, the stimuli were derived from videorecorded speeches. However, the mode of stimulus presentation differed between the two conditions. In the condition without SmarTerp, the recorded speech was played using VLC media player on the instructor's computer and mirrored to the participants' screens, allowing them to see and hear the speaker. In the SmarTerp condition, SmarTerp was run on the instructor's computer and mirrored to the participants' screens. That is, SmarTerp was playing the recorded speech and actively processing and extracting terms in real-time. This live interaction with SmarTerp was crucial to maintaining the authenticity of the tool's functionality during the study.

The participants were exposed to these two interfaces, but they did not interact with them since they were mirrored through the instructor's computer. Given this setup, all students in a specific data-collection session experienced identical stimuli with the same latency. The live prompting feature of SmarTerp was not captured on video for subsequent latency analysis. According to the tool's developers, the estimated latency ranges from a few milliseconds to a maximum of 1 second.

Our study is aptly classified as a quasi-experiment due to several defining characteristics inherent to its design. While we do exert control over the independent variable—determining whether participants interpret with or without SmarTerp—our study lacks a distinct control group and random assignment of participants to varied conditions or groups, both of which are pivotal components of a true experimental design. The absence of a proper control group and the non-random assignment inherently limit our ability to control for all potential confounders and, consequently, to definitively attribute observed differences to the manipulation of the independent variable. Additionally, the intra-subject design allows for comparisons within the same subjects under different conditions, but it does not compensate for the absence of random assignment and a proper control group. By clarifying that ours is a quasi-experiment, we acknowledge the constraints in making causal inferences and generalizations from our findings.

Before the first session, participants attended a self-training online course on the use of SmarTerp, so they could become familiar with the interface and its functionalities. In the first session, and before the first interpreting task, they were orally briefed about the study and could ask questions before signing the informed consent forms. Before each task, the participants were briefed on the text, the name of the speaker, and the event where the speech took place.

The use of Empatica E4 wristbands to measure physiological indicators (see § 2.4) called for adding some additional stages in the data-collection procedure, for it required recording a baseline. Before the initial interpreting task, participants were given roughly ten minutes to relax and to measure their baseline. After the second task, a similar relaxation period allowed for stress recovery measurement. However, recovery data were not used in the subsequent analysis (see below).

After performing the task in the two conditions, the participants filled in a questionnaire on the perceived usefulness of the tool. However, in the second data-collection session the participants only interpreted in one condition, with SmarTerp, and they did not fill in the questionnaire (Table 1). The purpose of this second session was to observe habituation effects (research question II) by facilitating an additional session wherein participants had further exposure to SmarTerp. Besides, in the first and third sessions, they were given a five-minute break between the two interpreting tasks. Table 1 summarizes the procedure.

	Before session 1	Session 1 (week 1)	Session 2 (week 2)	Session 3 (week 3)
Tasks	1. Webinar on SmarTerp	Q&A and informed consent		
		Putting wristbands on	Putting wristbands on	Putting wristbands on
		Ten-minute relaxation stage (baseline)	Ten-minute relaxation stage (baseline)	Ten-minute relaxation stage (baseline)
		Brief + interpreting task without SmarTerp		Brief + interpreting task without SmarTerp
		Five-minute pause		Five-minute pause
		Brief + interpreting task with SmarTerp	Brief + interpreting task with SmarTerp	Brief + interpreting task with SmarTerp
		Ten-minute relaxation stage (recovery)	Ten-minute relaxation stage (recovery)	Ten-minute relaxation stage (recovery)
		Questionnaire on SmarTerp		Questionnaire on SmarTerp
Time	1 h approx.	1 h 20 min. approx.	40 min. approx.	1 h 15 min. approx.

Table 1. Data-collection procedure

To be able to use the beta version of SmarTerp, developers had been provided with video recordings of the speeches, along with their respective glossaries containing potential problem triggers and their translations before each session (see §2.6). SmarTerp's speech recognition system was thus pre-trained for each speech, so as to ensure accurate identification and consistent prompting. Prior to the task, participants were not exposed to the glossary, so they were unaware of the pre-determined problem triggers and their translations. Product-oriented findings from this study (Russo et al. forthcoming) suggest increased accuracy when interpreting problem triggers in all conditions involving SmarTerp. We cannot be sure whether participants arrived at their renditions by themselves or by adopting the suggestions prompted by the tool.

2.3. Sample

Twelve second-year students of the MA in interpreting at the University of Bologna were recruited using a non-probabilistic, convenience sampling procedure. They were all female speakers of Italian as L1 and had a mean age of 24 (minimum 22, maximum 26). No participant suffered from heart-related issues, and none consumed alcoholic drinks at least in the 10 hours preceding the experiment, or caffeine at least one hour before data collection. Three participants were tobacco smokers (25%). None of them had any experience with using SmarTerp. The group was divided into four sub-groups of three students each, based on language combination and direction: (1) Italian→Spanish; (2) Spanish→Italian; (3) Italian→English, and (4) English→Italian. The data from participants interpreting into their L1/A language (English and Spanish into Italian) and for participants into their L2/B languages (Italian into English and Spanish) will be analyzed and reported separately.

2.4. Data-collection tool: Empatica E4 wristbands

The main benefit of Empatica E4 wristbands is that they are unobtrusive at data collection, enhancing ecological validity. Physiological data on stress were collected with the photoplethysmography sensor of Empatica E4 wristbands.[1] This sensor measures heart rate (HR), blood volume pulse (BVP),

[1] https://www.empatica.com/research/e4/, accessed 05/10/2022

and inter-beat interval (IBI) data. Empatica E4 wristbands have been deemed adequate for collecting heart rate (HR) and heart rate variability (HRV) data. McCarthy et al. (2016) compared the signal qualities of the Empatica E4 and the General Electric's SEER Light Extend Recorder holter portable electrocardiogram, a standard clinical device for atrial fibrillation detection. Their results showed that the E4 produced similar data quality to the holter device 85% of the time, with the holter performing better only 5% of the time. Schuurmans et al. (2020) tested the accuracy and predictive value of the Empatica E4 wristband against the Vrije University Ambulatory Monitoring System (VU-AMS) and found significant correlations between them for multiple HR and HRV indicators (see § 2.5). More research is needed, but these studies support using the E4 wristband for HR and HRV data collection. Overall, the Empatica E4 wristband has potential for measuring HR and HRV under non-movement conditions.

To minimize confounding effects due to stress induced by starting the study and by fatigue at the end of the task, only recordings of the same length were compared, following the guidelines of the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996). As recommended by Rojo & Korpal (2020), we extracted four 5-minute-long spans from the central minutes of each condition, following the rule of the three R's—Resting, Reactivity, and Recovery (Laborde et al. 2018)—i.e., baseline (resting, no task), interpreting without SmarTerp (task 1), interpreting with SmarTerp (task 2), and recovery (relaxation, after task 2). The baseline was measured with the participants sitting alone in their booths with both feet on the floor. Laborde et al. (2017) lists further recommendations for an optimal baseline—ankles at an angle of 90 degrees, hands on thighs, and eyes closed—but we thought that forcing the posture of participants in such detail would negatively affect the ecological validity. Again, following Laborde et al. (2017), Recovery was measured with the participants sitting in the same position as in the baseline one (Resting). Recovery data was not employed in this study (see § 2.5). Kubios HRV Premium software was used to automatically detect and correct artifacts, based on Lipponen & Tarvainen (2019). On average, 9.3% (Mdn = 9.4; SD = 2.0) of beats in each participant's recording were corrected.

2.5. Indicators

Before describing the indicators that we used to measure physiological stress, we need to frame and define this construct. The term arousal describes a state of physiological activation or cortical responsiveness, associated with sensory stimulation and activation of fibers from the reticular activating system. It may also refer to a state of excitement or energy expenditure linked to an emotion. Usually, arousal is closely related to a person's appraisal of the significance of an event or to the physical intensity of a stimulus. Arousal can either facilitate or debilitate performance. Responses can be positive (eustress), promoting focused attention and cognitive flow (Moneta 2020), or negative (distress), leading to anxiety or frustration (Kemeny 2003; but see Bienertova-Vasku et al. 2020 for a critique on the difference between these two notions). Stress is here understood as a stimulus, response, or physiological consequence, triggered by physical/environmental, task-related, and interpersonal stressors in interpreting (Cooper et al. 1982; Kemeny 2003). Stressful situations elicit physiological responses, including sympathetic nervous system activation, hypothalamic-pituitary-adrenal axis activation, and immunological cell level fluctuations (Kemeny 2003). However, repeated exposures to aversive stimuli or situations may lead people to adapt their physiological responses to stress. Adaptation here describes changes that reduce the physiological strain produced by stressful components of the total environment. Many organisms, not only humans, adapt to repeated stimuli by reducing their response, a feature described as habituation. With habituation we may refer here both to the process of growing accustomed to a situation or stimulus, or the diminished effectiveness of a stimulus in eliciting a response, following repeated exposure to the stimulus.[2]

Our study focused on sympathetic nervous system activation, measured through heart rate (HR) and heart rate variability (HRV). HR indicates the average number of heartbeats over a period, while HRV measures the variations in time between successive heartbeats. The indicators of stress based on HR and HRV employed in our study (see below) were computed from the BVP data measured with the wristbands. HR alone is mainly an indicator of physical exertion (Rojo & Korpal 2020), and it is often studied in the literature in conjunction with other HRV indicators or as a part of them (Kim et al. 2018). HRV is a complex measure, with different indicators providing information about the contribution of the autonomic nervous system to cardiac activity and regulation (Rojo & Korpal 2020: 196). That is, HRV accounts for the activity of the vagus nerve, thus "the focus is on vagal tone and its correlation with better executive cognitive performance, as well as better emotional and health regulation" (Rojo & Korpal 2020: 196). However, Rojo & Korpal also report that HRV is more sensitive to artifacts (for instance, induced by participants moving during data collection) and needs greater accuracy in measurement.

An increased HR mainly points at higher physical exertion, while an increased HRV—i.e., more variation between heartbeats over a certain period—would suggest a "greater ability to tolerate stress or [...] recovery from prior accumulated stress" (Rojo & Korpal 2020: 196), while a reduced variation between heartbeats suggests "stress from exercise, psychological events, or other internal or external stressors" (ibid.). Researchers have derived many indicators based on HR and HRV data, but not all of them are adequate to measure stress. Based on Rojo et al. (2021), on metanalyses (Castaldo et al. 2015; Schaffer & Ginsberg 2017; Kim et al. 2018) and on Spinolo et al. (2022), the HR and HRV indicators examined were the ones listed and defined in Table 2.

indicator	description (based on Schaffer et al. 2014 and Schaffer & Ginsberg 2017)
Mean HR	The **mean heart rate** associated to higher physical exertion is also expected to increase in a situation of stress, pointing at *sympathetic* activity, rather than to parasympathetic activity.
Mean RR	**R** is the peak point of successive R-peaks, or QRS complex, of the electrocardiogram wave. The mean of RR intervals is expected to decrease in a situation of stress, pointing again at sympathetic activity.
RMSSD	The **square Root of the Mean Squared Differences between Successive RR intervals.** It reflects the influence of parasympathetic activity on HRV. Under stress, RMSSD values tend to decrease, indicating a reduction in parasympathetic (vagal) activity and an increase in sympathetic activity, as part of the physiological response to stress.
LF-HF ratio	The ratio between **Low Frequency** and **High Frequency** band powers. A ratio above 1 indicates sympathetic activity (i.e., stress), while a value below 1 shows parasympathetic activity (i.e., relaxation).

Table 2. Heart rate and heart-rate variability indicators

Both Mean HR and Mean RR pertain to heart rate. So, they are related, but they differ in the aspects of heart rate they represent. Mean HR is the pulse rate, calculated by dividing the number of recorded heartbeats by the duration of that recording, in minutes. This yields and is expressed as an average, number of heartbeats per minute. Mean HR provides an overall estimate of the heart's activity level, but it does not convey information about heart rate variability. In contrast, Mean RR focuses on the time intervals between successive heartbeats. It refers to the average time interval between successive

[2] Definitions from the *APA Dictionary of Psychology*. For the physiology and neurobiology of stress and adaptation, see McEwen (2007). For workplace pressure at teleworking, see Day et al (2021) and Seeman et al (2023).

R-peaks (QRS complex) in an electrocardiogram recording, that is, the period between two heartbeats. This measure, typically expressed in milliseconds, captures the time span between heartbeats and informs on the variability of heart rate over time. By including both Mean RR and Mean HR, our study aimed to provide a comprehensive analysis to better understand the physiological responses under investigation.

In HRV research, baseline measurements work as reference points for the individual's autonomic function under neutral conditions, reflecting their inherent level of HRV (Veltman & Gaillard 1993; Moses et al. 2007; Loudon & Deininger 2016). The main method for evaluating the physiological responses elicited by a task is comparing baseline measurements with on-task indicators. The rationale lies in the ability to capture changes in autonomic nervous system (ANS) function and regulation that occur in response to the task, while accounting for individual differences in baseline HRV values. HRV indicators collected while at task are compared with the baseline measurements to assess the relative impact of the task on ANS activity—specifically, the balance between sympathetic (fight-or-flight) and parasympathetic (rest-and-digest) components. Changes in HRV indicators between the baseline and on-task conditions can reveal the extent to which the task induces stress or cognitive effort. To compare the baseline and on-task measurements of the four selected indicators (Table 2), on-task measurements were subtracted from the baseline values measured in that session. The results of this procedure indicate the **direction of change**, and are interpreted as follows:

- Mean HR increases with stress. If the difference between baseline and on-task measurement is positive, stress levels were higher at baseline measurement.
- Mean RR decreases with stress. If the difference between baseline and on-task measurement is negative, stress levels were higher at baseline measurement.
- RMSSD decreases with stress. If the difference between baseline and on-task measurement is negative, stress levels were higher at baseline measurement.
- A positive difference in LF/HF ratio points to a tendency towards parasympathetic activity (relaxation), while a negative difference shows a tendency towards sympathetic activity (stress).

To assess the **magnitude of difference** between baseline and on-task measurements, we calculated the median percent difference, which represents the central tendency of individual percent differences. Compared to the mean, the median is a more robust measure of central tendency. It is less susceptible to the influence of extreme values, thereby providing a more accurate representation of the data's center in limited sample sizes. The larger the median percent difference, the greater the difference between baseline and on-task measurement. The interpretation of the direction of the change follows the guidelines provided for each indicator.

2.6. Source texts as stimuli

Given the varied language-combinations involved in this study, the three data-collection sessions, and the two conditions (with and without SmarTerp), we needed to ensure that the main stimuli—the source speeches—were comparable in terms of topic, word count, text complexity, and delivery speed.

Each participant took part in five data-collection tasks, so they interpreted five speeches. The source languages were three: English, Italian, and Spanish. To reduce variation in source speeches, we created a source text for each session that was similar for the three source languages. To do so, one text was borrowed from Frittella (2023) and four texts with similar characteristics were crafted in English. The five texts were then translated into Spanish and Italian and proofread by L1 speakers of the respective languages. To further enhance comparability, all texts were of the same genre or type (opening speeches at international events), and they all had the same structure: an introduction, greetings, an agenda of the event, the body of the speech, and a conclusion.

We also calculated a readability measure for the five texts in each source language (English, Italian, and Spanish), using the Flesch-Kincaid index for English (Kincaid et al. 1975), GULPEASE index for Italian (Lucisano & Piemontese 1988), and Fernández's (1959) readability scale for Spanish.[3] The three indices have the same range (100 = very easy to read; 0 = very difficult to read). The English texts yielded a mean readability score of 49.9 (mdn = 47.9; SD = 5.9), suggesting *difficulty,* while Italian texts had a mean of 53 (mdn = 54; SD = 2.5) and Spanish texts a mean of 59.5 (mdn = 58.9; SD = 4.3), both suggesting *moderate difficulty.* The low data dispersion within each language reveals comparable readability levels, with texts ranging from moderately difficult to difficult across the three indices.

To ensure that the complexity of the five texts was similar, they were scanned for possible problem triggers, which were in turn categorized based on Frittella (2023). The aim was to obtain well-balanced texts that contained parallel potential problem triggers. For each language pair, SmarTerp was fed with a glossary with the corresponding specialized terms, proper names (entities and persons), and numbers in the appropriate language pair, so that the output offered by the tool would be the same for all participants. Of course, the difficulty of the texts may have been perceived differently for each participant, due to their personal characteristics.

The five speeches were videorecorded by L1 speakers. The average delivery rate for the five speeches in English was 107.4 words per minute (mdn = 108.3; SD = 9.7), 101.7 words per minute for the Italian speeches (mdn = 101.7; SD = 3.0), and 104.8 for the Spanish speeches (mdn = 106.1; SD = 5.9). The mean duration of the recordings of the English speeches was 09:48 minutes (mdn = 09:55; SD = 00:31), 10:45 minutes for the Italian speeches (mdn = 10:47; SD = 00:43), and 10:54 for the Spanish speeches (mdn = 10:55; SD = 01:02).

2.7. Statistical analysis

Given the intra-subject nature of our design, a single datapoint per (1) participant, (2) measurement (baseline, interpreting without SmarTerp, interpreting with SmarTerp), and (3) directionality—into the A or B language—makes a sample size of N = 6 too small to apply statistical procedures such as regression analyses. Language pair may have had an impact on the participants' stress levels, yet we merged all participants interpreting into their A language in one group and those interpreting into their B language in another group, since the facilitation effect of the tool and subsequent reduction in stress should be equally present in all cases.

Log transforming HR and HRV indicators to normalize the data (as recommended by Laborde et al. 2017) would have had little impact on the analysis with such limited sample size. We thus employed the non-parametric procedures described in Table 3.

The significance level was pre-established at α = 0.05. To establish a minimum effect size of interest, we performed a sensitive analysis of the Wilcoxon signed-rank test for matched pairs for a unilateral test with α = 0.05, power = 0.8, and a sample size of 6 using G*Power 3.1.9.6. The minimum effect size of interest is $d = 1.226$. Converted into r (Ruscio 2008), it equals 0.522. Hence, effect sizes under 0.522 were considered irrelevant even if statistically significant. Statistical analysis was carried out in RStudio Desktop and jamovi 2.3 (jamovi project 2023).

[3] Readability formulas are atheoretical and unreliable, but there is no consensus in the field as to how to objectively profile texts and quantify relevant text features. However, the indicators combined into these formulas (e.g., word frequency, sentence length) do have supporting evidence of their merit in isolation. CTIS research has often used readability formulas for text profiling, so these results are offered merely as a reference, whereas the comparability between the texts was rather achieved by the other steps. The differences between English and the romance language translations might be the consequence of having been processed and revised by professional communicators.

Aim	Test	Contrast type	Effect size
Comparing within-group baseline vs. on-task measurement for a given indicator in a given session	Wilcoxon signed-rank test with Bonferroni correction	Unilateral	*r*
Comparing within-group median percent difference for a given indicator between two conditions	Wilcoxon signed-rank test with Bonferroni correction	Unilateral	*r*
Comparing between-group median percent difference for a given indicator in a given session and condition	Mann-Whitney *U* test	Unilateral	*r*
Comparing within-group median percent difference for a given indicator among the three sessions in the with-SmarTerp condition	Friedman test as omnibus test and Wilcoxon signed-rank test with Bonferroni correction as post-hoc test	Omnibus: bilateral; post-hoc: unilateral	*W*
Comparing within-group median percent difference for a given indicator between the two sessions in the no-SmarTerp condition	Wilcoxon signed-rank test with Bonferroni correction	Unilateral	*r*

Table 3. Non-parametric statistical procedures

3. Results and discussion

Table 4 presents the descriptive results of the difference between the baseline and the on-task measurement of each direction (into A and into B language), indicator (Mean HR, Mean RR, RMSSD, and LF/HF ratio), and condition (with and without SmarTerp). It also provides the median percent difference between baseline and on-task measurement.

Difference			Session 1				Session 2				Session 3			
(Baseline – on-task measurement)			Mean HR	Mean RR	RMSSD	LF/HF ratio	Mean HR	Mean RR	RMSSD	LF/HF ratio	Mean HR	Mean RR	RMSSD	LF/HF ratio
A lang.	With	Median	−4.0	57.0	28.0	−0.1	0.5	−5.5	18.5	0.2	−3.0	35.0	94.0	-0.1
		SD	21.4	217.2	160.6	1.3	17.1	201.1	136.7	1.1	9.4	124.4	176.5	1.7
		Median % diff.	*−6.4*	*6.0*	*6.8*	*−10.0*	*0.8*	*−4.3*	*−0.6*	*9.1*	*−4.0*	*4.3*	*48.8*	*-3.6*
	Without	Median	−5.5	79.5	68.5	0.0					−6.5	62.5	73.2	−0.6
		SD	23.7	245.3	230.0	0.9		-			7.8	109.4	161.7	1.3
		Median % diff.	*−8.8*	*8.3*	*16.4*	*4.8*					*−8.1*	*8.2*	*41.7*	*−24.1*
B lang.	With	Median	5.0	−68.0	−45.7	0.0	−4.0	54.5	27.9	0.1	−9.0	91.0	124.0	-0.4
		SD	11.7	123.3	113.1	0.4	8.0	102.1	83.8	0.7	10.1	104.8	123.2	1.8
		Median % diff.	*7.5*	*−7.6*	*−11.3*	*−1.0*	*−6.2*	*6.4*	*6.1*	*7.4*	*−12.0*	*12.0*	*44.7*	*-47.0*
	Without	Median	−19.0	187.5	177.5	0.2					−8.0	83.5	55.8	−1.9
		SD	17.2	167.5	136.8	1.3		-			10.6	110.1	154.8	2.2
		Median % diff.	*−25.5*	*25.6*	*58.2*	*14.5*					*−10.3*	*10.5*	*39.3*	*−57.1*

Table 4. Descriptive statistics of the difference between baseline and on-task measurement for each indicator, direction, and condition

Interpretation of results is as follows:

- Mean HR increases with stress. If baseline − on-task measurement is positive, stress levels were higher during baseline.
- Mean RR decreases with stress. If baseline − on-task measurement is negative, stress levels were higher during baseline.
- RMSSD decreases with stress. If baseline − on-task measurement is negative, stress levels were higher during baseline.
- LF/HF ratio. Positive difference → tendency towards parasympathetic activity (relaxation). Negative difference → tendency towards sympathetic activity (stress).
- Median percent difference indicates the size of the difference between the two measurements (the larger the value, the larger the difference). The interpretation of the direction of the change follows the guidelines provided for each indicator.

3.1. Stress levels with vs. without SmarTerp

3.1.1. Interpreting into the A language (L1)

In the results from participants interpreting into their A language using SmarTerp in session 1, the median values for Mean HR (−4.0) and Mean RR (57.0) suggest increased stress levels in the on-task measurement compared to the baseline. Concurrently, RMSSD and LF/HF ratio exhibit the same directional changes (28.0 for RMSSD; −0.1 for LF/HF ratio). However, inferential tests reveal no significant differences among the four indicators between measurements. They support that, in session 1, interpreting with SmarTerp into the A language did not yield lower stress levels than the baseline (Table 5).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	11.0	1.000	0.048
Mean RR	10.0	1.000	−0.048
RMSSD	10.0	1.000	−0.048
LF/HF ratio	7.0	1.000	−0.333

Table 5. Results of the Wilcoxon signed-rank test for the with-SmarTerp condition in session 1 for participants interpreting into their A language. Note: all tests are unilateral

Also in session 1, when interpreting into the participants' A language without SmarTerp, the median values of Mean HR (−5.5), Mean RR (79.5), and RMSSD (68.5) show a consistent direction of change between baseline and on-task measurements. The three indicators suggest increased stress levels during task performance, compared to the baseline. However, the LF/HF ratio exhibits no variation between measurements (0.0). Despite these observations, inferential analysis reveals no significant differences in stress levels between the baseline and interpreting without SmarTerp (Table 6).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	5.0	1.000	−0.333
Mean RR	14.0	1.000	0.333
RMSSD	15.0	1.000	0.429
LF/HF ratio	10.0	1.000	−0.048

Table 6. Results of the Wilcoxon signed-rank test for the no-SmarTerp condition in session 1 for participants interpreting into their A language. Note: all tests are unilateral

Elevated stress levels when at task compared to the baseline were anticipated, given the inherently stressful nature of interpreting, irrespective of SmarTerp usage. The relative reduction in stress levels with SmarTerp, compared to the task without the tool, is assessed by comparing median percent differences. Comparing this indicator between the with- and no-SmarTerp conditions for participants interpreting into their A language in session 1, a larger magnitude of difference between baseline and on-task measurements is observed in the without condition for Mean HR (−8.8 vs. −6.4), Mean RR (8.3 vs. 6.0), and RMSSD (16.4 vs. 6.8). Conversely, LF/HF ratio yields an opposite trend (without 4.8 vs. with −10). However, these differences are not statistically significant (Table 7), suggesting that interpreting with or without SmarTerp in session 1 does not impact the magnitude of difference between baseline and on-task measurements for participants' A language.

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	16.0	1.000	0.524
Mean RR	6.0	0.876	0.429
RMSSD	7.0	1.000	0.333
LF/HF ratio	6.0	1.000	0.200

Table 7. Wilcoxon signed-rank test results comparing median percent difference in session 1 under both SmarTerp conditions for participants interpreting into their A language. Note: all tests are unilateral

Session 2 results for participants using SmarTerp indicate an opposite trend compared to session 1. The median values for Mean HR (0.5), Mean RR (−5.5), and LF/HF ratio (0.2) suggest elevated stress levels in the baseline compared to on-task measurements, while the median value for RMSSD (18.5) suggests the opposing trend. No significant differences were identified between baseline and on-task measurements (Table 8).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	8.0	1.000	0.067
Mean RR	9.0	1.000	−0.143
RMSSD	12.0	1.000	0.143
LF/HF ratio	10.0	1.000	−0.048

Table 8. Results of the Wilcoxon signed-rank test for the no-SmarTerp condition in session 2 for participants interpreting into their A language. Note: all tests are unilateral

In session 3, participants exhibit increased stress levels in the on-task measurement, relative to the baseline for three of the four indicators in the with-SmarTerp condition: Mean HR (−3.0), Mean RR (35.0), and RMSSD (94.0). The direction of change for LF/HF ratio suggests increased sympathetic activity when interpreting, albeit with a minimal magnitude of difference (−0.1). Nonetheless, inferential tests do not identify significant differences between baseline and on-task measurements (Table 9).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	5.5	1.000	−0.476
Mean RR	15.0	1.000	0.429
RMSSD	17.0	1.000	0.619
LF/HF ratio	8.0	1.000	−0.238

Table 9. Results of the Wilcoxon signed-rank test for the with-SmarTerp condition in session 3 for participants interpreting into their A language. Note: all tests are unilateral

In session 3, participants interpreting into their A language without SmarTerp display increased stress levels in on-task measurements, compared to the baseline for three of the four indicators: Mean HR (−6.5), Mean RR (62.5), and RMSSD (73.2). The LF/HF ratio presents a direction of change indicating sympathetic activity (−0.6). As in prior cases, inferential analysis reveals no significant differences between baseline and on-task measurements (Table 10).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	4.0	1.000	−0.619
Mean RR	17.0	1.000	0.619
RMSSD	16.0	1.000	0.524
LF/HF ratio	9.0	1.000	−0.143

Table 10. Results of the Wilcoxon signed-rank test for the no-SmarTerp condition in session 3 for participants interpreting into their A language. Note: all tests are unilateral

Comparing the median percent difference between with- and no-SmarTerp conditions in session 3, the no-SmarTerp condition shows a larger magnitude of difference between baseline and on-task measurements for all indicators except RMSSD: Mean HR (with: −4.0; without: −8.1), Mean RR (with: 4.3; without: 8.2), RMSSD (with: 48.8; without: 41.7), LF/HF ratio (with: −3.6; without: −24.1). This suggests that, although participants experienced increased stress levels during tasks in both conditions, these levels were higher without SmarTerp relative to the baseline compared to the with-SmarTerp condition. Nevertheless, inferential statistics reveal no significant differences (Table 11).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	8.0	1.000	0.067
Mean RR	11.0	1.000	0.048
RMSSD	10.0	1.000	0.048
LF/HF ratio	11.0	1.000	0.048

Table 11. Wilcoxon signed-rank test results comparing median percent difference in session 3 under both SmarTerp conditions for participants interpreting into their A language. Note: all tests are unilateral

To sum up, from a descriptive standpoint, sessions 1 and 3 exhibited elevated stress levels in A-language interpreting, both with and without SmarTerp, relative to the baseline. The magnitude of this difference was more pronounced in the no-SmarTerp condition in both sessions, potentially signaling the stress-mitigating assistance of SmarTerp. Conversely, session 2 displayed contrasting trends, with the majority of indicators suggesting higher stress levels at the baseline than during the on-task measurements. This anomaly might be effects of extraneous variables.

3.1.2. Interpreting into the B language (L2)

In session 1, participants interpreting with SmarTerp into their B language exhibited lower stress levels during tasks compared to the baseline for Mean HR (5.0), Mean RR (−68.0), and RMSSD (−45.7). The LF/HF ratio demonstrated no change, with a median of 0.0. However, inferential statistics reveal no significant differences between on-task measurements and the baseline (Table 12).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	16.0	0.624	0.524
Mean RR	5.0	0.624	−0.524
RMSSD	8.0	1.000	−0.238
LF/HF ratio	8.0	1.000	−0.238

Table 12. Results of the Wilcoxon signed-rank test for the with-SmarTerp condition in session 1 for participants interpreting into their B language. Note: all tests are unilateral

The no-SmarTerp condition exhibited an opposite trend. Participants seemed to have increased stress levels during tasks compared to the baseline: Mean HR (−19.0), Mean RR (187.5), and RMSSD (177.5). The median (0.2) of the LF/HF ratio displays a minimally divergent trend. Inferential statistics reveal no significant differences (Table 13).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	3.0	1.000	−0.714
Mean RR	18.0	1.000	0.714
RMSSD	18.0	1.000	0.714
LF/HF ratio	10.0	1.000	−0.048

Table 13. Results of the Wilcoxon signed-rank test for the no-SmarTerp condition in session 1 for participants interpreting into their B language. Note: all tests are unilateral

The median percent difference in session 1 reveals a larger magnitude of difference in the no-SmarTerp condition for all indicators: Mean HR (with: 7.5; without: −25.5), Mean RR (with: −7.6; without: 25.6), RMSSD (with: −11.3; without: 58.2), LF/HF ratio (with: −1.0; without: 14.5). Inferential results are nonsignificant (Table 14), but descriptive findings for session 1 indicate decreased stress levels, compared to the baseline, when participants interpret into their B language using SmarTerp; increased levels, without it; and a substantially larger difference between the baseline and on-task measurements in the no-SmarTerp condition. This contrasts with results for participants interpreting into their A language in session 1, where stress levels were not reduced with SmarTerp and the magnitude of difference between baseline and on-task measurement was less pronounced.

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	21.0	1.000	1.000
Mean RR	0.0	0.064	1.000
RMSSD	0.0	0.064	1.000
LF/HF ratio	12.0	1.000	0.143

Table 14. Wilcoxon signed-rank test results comparing median percent difference in session 1 under both SmarTerp conditions for participants interpreting into their B language. Note: all tests are unilateral

In contrast to session 1, session 2 results exhibit increased stress levels during task performance compared to the baseline for Mean HR (−4.0), Mean RR (54.5), and RMSSD (27.9). The LF/HF ratio (0.1) indicates almost no change. Nevertheless, differences between baseline and on-task measurements were not significant (Table 15).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	5.0	1.000	−0.524
Mean RR	16.0	1.000	0.524
RMSSD	20.0	1.000	0.905
LF/HF ratio	12.0	1.000	0.143

Table 15. Results of the Wilcoxon signed-rank test for the no-SmarTerp condition in session 2 for participants interpreting into their B language. Note: all tests are unilateral

Session 3 with SmarTerp exhibits a trend contrary to that in session 1 and follows the one detected in session 2. Participants display increased stress during task performance, compared to the baseline: Mean HR (−9.0), Mean RR (91.0), and RMSSD (124.0). The LF/HF ratio also indicates sympathetic activity (−0.4). Inferential results reveal no significant differences between baseline and on-task measurements (Table 16).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	5.5	1.000	−0.476
Mean RR	14.0	1.000	0.333
RMSSD	17.0	1.000	0.619
LF/HF ratio	8.0	1.000	−0.238

Table 16. Results of the Wilcoxon signed-rank test for the with-SmarTerp condition in session 3 for participants interpreting into their B language. Note: all tests are unilateral

In session 3, participants interpreting without SmarTerp exhibit the same trend as with the tool, with increased stress levels while at task, compared to the baseline: Mean HR (−9.0), Mean RR (91.0), RMSSD (124.0), and LF/HF ratio (−1.9). Inferential analysis again reveals no significant differences between the baseline and on-task measurements (Table 17).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	4.5	1.000	−0.571
Mean RR	17.0	1.000	0.619
RMSSD	17.0	1.000	0.619
LF/HF ratio	3.0	1.000	−0.714

Table 17. Results of the Wilcoxon signed-rank test for the no-SmarTerp condition in session 3 for participants interpreting into their B language. Note: all tests are unilateral

The median percent difference in session 3 reveals that, except for the LF/HF ratio (with: −47.0; without: −57.1), the remaining indicators exhibit a larger magnitude of difference between baseline and on-task measurements in the with-SmarTerp condition: Mean HR (with: −12.0; without: −10.3), Mean RR (with: 12.0; without: 10.5), and RMSSD (with: 44.7; without: 39.3). Nonetheless, these differences are not statistically significant (Table 18).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	14.0	1.000	0.333
Mean RR	7.0	1.000	0.333
RMSSD	7.0	1.000	0.333
LF/HF ratio	17.0	1.000	0.619

Table 18. Wilcoxon signed-rank test results comparing median percent difference in session 3 under both SmarTerp conditions for participants interpreting into their B language. Note: all tests are unilateral.

Session	Condition	Median percent difference for...	A language	B language	*U*	*r*
1	With	Mean HR	−6.4	7.5	16.0	0.111
		Mean RR	6.0	−7.6	16.0	0.111
		RMSSD	6.8	−11.3	16.0	0.111
		LF/HF ratio	−10.0	−1.0	17.0	0.056
1	Without	Mean HR	−8.8	−25.5	13.0	0.278
		Mean RR	8.3	25.6	13.0	0.278
		RMSSD	16.4	58.2	14.0	0.222
		LF/HF ratio	4.8	14.5	17.0	0.056
2	With	Mean HR	0.8	−6.2	15.0	0.167
		Mean RR	−4.3	6.4	15.0	0.167
		RMSSD	−0.6	6.1	14.0	0.222
		LF/HF ratio	9.1	7.4	18.0	0.000
3	With	Mean HR	−4.0	−12.0	16.0	0.111
		Mean RR	4.3	12.0	16.0	0.111
		RMSSD	48.8	44.7	15.0	0.167
		LF/HF ratio	−3.6	−47.0	14.0	0.222
3	Without	Mean HR	−8.1	−10.3	17.0	0.056
		Mean RR	8.2	10.5	17.0	0.056
		RMSSD	41.7	39.3	17.0	0.056
		LF/HF ratio	−24.1	−57.1	12.0	0.333

Table 19. Mann-Whitney *U* test results comparing median percent difference between participants interpreting into their A and B languages for each session, condition, and indicator. Note: all tests are unilateral. All *p*-values were non-significant at $p = 1.000$.

In short, the descriptive analysis indicates reduced stress levels when interpreting into the B language with SmarTerp only in the first session. The stress difference between baseline and task performance was notably larger without SmarTerp. Although interpreting with SmarTerp remains stressful, it is less so than without it. Nonetheless, uncontrolled confounding variables may influence these results.

When comparing the results of participants interpreting into their A and B languages as independent groups, the median percent differences are generally larger when interpreting into B language in both conditions and all sessions, albeit nonsignificant (Table 19). This result may suggest that interpreting with SmarTerp has a stronger stress-reducing effect when done into the L2.

3.2. Stress levels with and without SmarTerp over time

3.2.1. Interpreting into the A language (L1)

In interpreting into the A language using SmarTerp, descriptive results reveal no consistent reduction in stress levels over time. In Session 1, most indicators suggested increased stress levels during the task, compared to the baseline (Mean HR: −4.0; Mean RR: 57.0; RMSSD: 28.0; LF/HF: −0.1). Session 2 exhibited a contrasting pattern, with Mean HR (0.5), Mean RR (−5.5), and LF/HF ratio (0.2) displaying increased stress levels during the baseline, while only RMSSD demonstrated increased stress levels when at task (18.5). In Session 3, stress levels were noticeably higher during the task than at the baseline for all indicators (Mean HR: −3.0; Mean RR: 35.0; RMSSD: 94.0; LF/HF ratio: −0.1). However, due to the lack of task randomization, it is unclear if the increase in stress in session 3 resulted from tool usage, increased text difficulty perception, poor performance, or a combination thereof. The omnibus Friedman test for median percent difference across the three sessions produced no significant findings (Table 20).

Median percent difference for…	χ^2	*p*	*W*
Mean HR	0.0	1.000	0.000
Mean RR	0.0	1.000	0.000
RMSSD	0.3	0.846	0.028
LF/HF ratio	0.3	0.846	0.028

Table 20. Friedman test results comparing median percent difference for the three sessions for participants interpreting into their A language with SmarTerp. Note: degrees of freedom for all tests = 2

As for interpreting without SmarTerp into the A language, there was a consistent trend of increased stress levels during the task, compared to the baseline, was observed in both session 1 (Mean HR: −5.5; Mean RR: 79.5; RMSSD: 68.5; LF/HF ratio: 0.0) and session 3 (Mean HR: −6.5; Mean RR: 62.5; RMSSD: 73.2; LF/HF ratio: −0.6). The median percent difference in magnitude was comparable for two indicators (Mean HR in session 1: −8.8; session 3: −8.1; Mean RR in session 1: 8.3; session 3: 8.2), while for RMSSD (session 1: 16.4; session 3: 41.7) and LF/HF ratio (session 1: 4.8; session 3: −24.1), the stress level differences were substantially greater in the third session. Inferential statistics revealed no significant differences between Session 1 and 3 in terms of median percent difference (Table 21).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	9.0	1.000	−0.143
Mean RR	11.0	1.000	0.048
RMSSD	7.0	1.000	0.333
LF/HF ratio	10.0	1.000	−0.048

Table 21. Results of the Wilcoxon signed-rank test for the no-SmarTerp condition in session 1 and 3 for participants interpreting into their A language. Note: all tests are unilateral

3.2.2. Interpreting into the B language

When interpreting into the B language, stress levels when using SmarTerp were lower, compared to the baseline, but only in session 1 (Mean HR: 5.0; Mean RR: −68.0; RMSSD: −45.7). Session 2 (Mean HR: −4.0; Mean RR: 54.5; RMSSD: 27.9) and session 3 (Mean HR: −9.0; Mean RR:

91.0; RMSSD m: 124.0) revealed increased stress levels when at task, compared to the baseline. The LF/HF ratio indicated negligible change in sessions 1 (0.0) and 2 (0.1), while in session 3, it suggested sympathetic activity (−0.4). As for the magnitude of median percent difference, session 3 displayed a higher stress level increase (Mean HR: −12.0; Mean RR: 12.0; RMSSD: 44.7; LF/HF ratio: −47.0) than session 2 (Mean HR: −1.0; Mean RR: −6.2; RMSSD: 6.4; LF/HF ratio: 7.4). Nonetheless, inferential statistics did not reveal significant differences between the three sessions (Table 22).

Median percent difference for…	χ^2	*p*	*w*
Mean HR	2.3	0.311	0.194
Mean RR	2.3	0.311	0.194
RMSSD	3.0	0.223	0.250
LF/HF ratio	1.0	0.607	0.083

Table 22. Friedman test results comparing median percent difference for the three sessions for participants interpreting into their B language with SmarTerp. Note: degrees of freedom for all tests = 2

Interpreting into the B language without SmarTerp revealed increased stress levels during tasks compared to the baseline in sessions 1 (Mean HR: −19.0; Mean RR: 187.5; RMSSD: 177.5; LF/HF ratio: 0.2) and 3 (Mean HR: −8.0; Mean RR: 83.5; RMSSD: 55.8; LF/HF ratio: −1.9). However, the magnitude of median percent difference was higher in session 1 (Mean HR: −25.5; Mean RR: 25.6; RMSSD: 58.2; LF/HF ratio: 14.5) compared to session 3 (Mean HR: −10.3; Mean RR: 10.5; RMSSD: 39.3; LF/HF ratio: −57.1). This suggests a decrease in stress levels in the final session, but there was no statistical difference between the two sessions (Table 23).

Indicator	*W*	*p* (Bonferroni)	*r*
Mean HR	6.0	0.876	−0.429
Mean RR	15.0	1.000	0.429
RMSSD	12.0	1.000	0.143
LF/HF ratio	18.0	1.000	0.714

Table 23. Results of the Wilcoxon signed-rank test for the no-SmarTerp condition in session 1 and 3 for participants interpreting into their B language. Note: all tests are unilateral

In brief, stress when using SmarTerp does not seem to decrease with exposure to the tool, regardless of whether participants interpret into their A or B languages. Perhaps the exposure period was too short for the participants to really be able to adapt their behavior to the features of the software, which might, as explained, require increases in their ear-voice span. Novice interpreters might need more time to adapt their ways than experienced interpreters, and they might have a harder time when faced with multiple sources of input (audio, video, written) when facing additional instances of decision-making to their interpreting (e.g., ‘shall I use the suggested term?’), and additional attention-drawing areas on the screen. Moreover, the speeches used in the study were different from each other. Their order was not randomized, which might have influenced how stressful a session was, whether using SmarTerp or not. Follow-up research comparing novices and experts, randomization of texts and tasks, and a design that includes a longer and more intense use of the tool should shed light on this issue.

Similarly, stress does not change between sessions 1 and 3 when the tool is not used. This, again, might be due to text and order effects, as interpreting without the tool was always the first task in a session. However, interpreting without SmarTerp was more stressful in all cases, which may indicate that the use of SmarTerp slightly reduces stress levels.

Finally, even if SmarTerp did not appear to reduce stress while interpreting, user experience and cost-benefit considerations should be taken on board as well. It is unclear whether this increased stress is perceived as relevant by participants. Self-reported measures, such as those obtained through validated psychometric tests, focus groups, and interviews would be useful to understand the effects of speech and speech order effects.

The focus of attention may also modulate the emotional response and influence the physiological response to potentially stressful situations. Wadlinger & Isaacowitz (2011) argue that attention focus on a stimulus can also regulate emotions by suppressing the processing of irrelevant stimuli. Translating is taken to be a more demanding task than reading for information. For instance, Rojo & Naranjo (2021) show that negative affect and stress levels of translation students were higher after simply reading an emotionally charged text than after translating it. Other studies, such as Rojo, Cifuentes & Espín (2021), Rojo, Foulquié, Espín & Martínez (2021), and Rojo, Cifuentes & López (2021) have shown that attention focus while translating can also modulate the participants' physiological response to emotional stress (e.g., cortisol or heart rate).

4. Concluding remarks

This paper has presented the results of a quasi-experimental, intra-subject study on the effects of the use of SmarTerp on physiological stress levels in simultaneous interpreting tasks carried out by MA students in interpreting with the aim to answer (I) whether it is less stressful to interpret with SmarTerp for such students, and (II) whether exposure to SmarTerp decreases stress over time.

The descriptive results suggest elevated stress levels when using SmarTerp in two out of the three sessions, compared to their baselines. This was anticipated, considering the inherently stressful nature of interpreting tasks, especially for students. The stress levels during task execution were predicted to surpass those in a relaxed state, irrespective of SmarTerp use. However, these descriptive observations did not translate into significant inferential outcomes: all inferential statistical results were non-significant.

Another descriptive trend was higher stress levels in the no-SmarTerp condition compared to using it, hinting at a potential stress-alleviating effect of the tool. This descriptive trend was discernible in both B→A (into L1) and A→B (into L2) interpreting, but the effect was more pronounced when interpreting into the B language. This could potentially be attributed to SmarTerp aiding in facing terminological units, proper names, and numbers, allowing students to focus more on linguistic quality during their performance. However, these descriptive trends were not substantiated by inferential statistics, which consistently revealed no significant differences.

One may ponder, then, how useful such a tool can be, in view that such stress reduction is non-significant and that, as discussed, the tool may lead to higher cognitive effort due to more complex decision-making and information processing. The answer to this question is not straightforward due to several impacting factors.

First, there can be many reasons for the interpreter to feel stress. For instance, the environment where data-collection sessions were conducted might have induced stress in the participants, or the elements SmarTerp provides assistance with might not have been stress triggers, whereas the topic, the delivery speed, or the speaker's diction were. Second, the use of SmarTerp could be adding extra layers of decision-making in the interpreting process (i.e., considering whether the solution provided is necessary, adequate, accurate, etc.), thereby increasing stress.

Third, even if stress levels are not lowered with SmarTerp to a significant extent, the tool may serve its purpose of aiding the interpreter in accurately rendering terminological units, proper names, and numbers. The project to which this study belongs also aims at investigating the accuracy of the participants' renditions of such units when interpreting with and without

SmarTerp, and if and how the number of accurate renditions correlates with stress. These results will be reported elsewhere.

Both descriptive and inferential statistics showed no decrease in stress levels when using SmarTerp over time, with various potential explanations. Apart from discussed factors, the single-month period encompassing three interpreting sessions with SmarTerp may have been insufficient to create habituation effects, thus preventing the observation of stress reduction related to tool usage.

This study has several limitations, so the results should be interpreted with caution. A first, obvious limitation is sample size (N = 12), which inevitably affected the power of statistical tests. A second limitation is that texts and task order were not randomized. As for stress analysis, no self-reported measures of stress were collected through questionnaires, interviews or focus groups to cross-reference results with those obtained with Empatica E4 wristbands through the analysis of HRV as a physiological indicator of stress. Finally, the speeches had to be necessarily different through the tasks and sessions. While they were controlled for several features that may affect performance and stress levels, other potentially relevant features could not be controlled, such as each participant's subjective perception of task difficulty.

In addition to correlating stress levels with the participants' renditions, an analysis which is underway, the results of this study suggest further scopes for follow-up studies. For instance, our participants were at the same stage of their training, so they may have had comparable levels of interpreting expertise. The use of SmarTerp at different stages of training could be studied to determine whether it has positive effects on reducing stress levels. A longitudinal study would also allow us to observe whether student participants adapt their interpreting strategies to the platform. A further line of work might compare the stress levels of interpreting students with those of professional interpreters when using SmarTerp while at task.

Funding and conflict of interest

This research was funded by EIT Digital within the SmarTerp project (ref. number 21184). Although the software will be commercialized, the authors had no conflict of interest, for they only participated in the piloting stage of the product and will not profit from its commercialization.

References

Biagini, G. (2016). *Glossario cartaceo e glossario elettronico durante l'interpretazione simultanea: uno studio comparativo.* MA thesis. Trieste: SSLiMIT Trieste.

Bienertova-Vasku, J., Lenart, P. & Scheringer, M. (2020). *Eustress and distress: Neither good nor bad, but rather the same?* In *BioEssays* 42(7), 900238.

Braun, S. (2019). Technology and interpreting. In O'Hagan, M. (eds.), *Routledge Handbook of Translation and Technology* (271–272). London: Routledge,

Castaldo, R., Melillo, P., Bracale, U., Caserta, M., Triassi, M. & Pecchia, L. (2015). Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. In *Biomedical Signal Processing and Control* 18, 370–377.

Chmiel, A. & Spinolo, N. (2022). *Testing the impact of remote interpreting settings on interpreter experience and performance: Methodological challenges inside the virtual booth.* In *Translation, Cognition & Behavior* 5(2), 250–274.

Cooper, C. L., Davies, R. & Tung, R. L. (1982). *Interpreting stress: Sources of job stress among conference interpreters.* In *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* 1(2), 97–108.

Day, A., Cook, R., Jones-Chick, R. & Myers, V. (2021). Are your smart technologies "killing it" or killing you? Developing a research agenda for workplace ICT and worker wellbeing. In Kelloway, E. K. & Cooper, C. (eds.), *A research agenda for workplace stress and wellbeing* (91–118). Berlin: Springer,

Defrancq, B. & Fantinuoli, C. (2021). *Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers.* In *Target. International Journal of Translation Studies* 33(1), 73–102.

Desmet, B., Vandierendonck, M. & Defranq, B. (2018). Simultaneous interpretation of numbers and the impact of technological support. In Fantinuoli, C. (eds.), *Interpreting and Technology* (13–27). Berlin: Language Science

Press,

Díaz-Galaz, S. (2015). *La influencia del conocimiento previo en la interpretación simultánea de discursos especializados: un estudio empírico*. PhD thesis. Granada: Universidad de Granada.

Fantinuoli, C. (2018). Interpreting and technology: The upcoming technological turn. In Fantinuoli, C. (eds.), *Interpreting and Technology* (1–12). Berlin: Language Science Press,

Fantinuoli, C. (2018). Computer-assisted interpretation: Challenges and future perspectives. In Corpas Pastor, G. & Durán-Muñoz, I. (eds.), *Trends in E-Tools and Resources for Translators and Interpreters* (153–174). Leiden: Brill.

Faul, F., Erdfelder, E., Buchner, A. & Lang, Albert-Georg (2009). *Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses.* In *Behavior Research Methods* 41, 1149–1160.

Fernández-Huerta, J. (1959). *Medidas sencillas de lecturabilidad*. In: *Consigna* 214, 29-32.

Frittella, Francesca M. 2022: CAI tool-supported SI of numbers: A theoretical and methodological contribution. In *International Journal of Interpreter Education* 14(1), 32–56.

Frittella, F. M. (2023). *Usability research for interpreter-centred technology: The case study of SmarTerp*. Berlin: Language Science Press.

Gacek, M. (2015). *Softwarelösungen für DolmetscherInnen*. MA thesis. Wien: Universität Wien.

Gile, D. (200)9). *Basic Concepts and Models for Interpreter and Translator Training: Revised Edition*. Amsterdam, John Benjamins.

Jamovi project (2023). J*amovi* (Version 2.3) [Computer Software]. https://www.jamovi.org

Kemeny, M. E. (2003). *The psychobiology of stress*. In *Current Directions in Psychological Science* 12(4), 124–129.

Kim, H., Cheon, E., Bai, D. Lee, Y. H. & Koo, B. (2018). *Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry Investigation* 15(3), 235–245.

Kincaid, J. P., Fishburne, R. P., Rodgers, R. L. & Chisomm, B. S. (1975). *Derivation of new readability formulas for Navy enlisted personnel*. Millington, TN: Chief of Naval Training.

Laborde, S., Mosley, E. & Thayer, J. F. (2017). *Heart rate variability and cardiac vagal tone in psychophysiological research – Recommendations for experiment planning, data analysis, and data reporting.* In *Front. Psychol*. 8:213.

Laborde S., Mosley, E. & Mertgen, A. (2018). *Vagal tank theory: The three Rs of cardiac vagal control functioning - Resting, Reactivity, and Recovery.* In *Front Neurosci.* 12:458.

Lipponen, J. A. & Tarvainen, M. P. (2019). *A robust algorithm for heart rate variability time series artefact correction using novel beat classification*. In *Journal of Medical Engineering & Technology* 43(3), 173–181.

Loudon, G. H. & Deininger, G. M. (2016). *The physiological response during divergent thinking.* In *Journal of Behavioral and Brain Science* 6(1), 28–37.

Lucisano, P. & Piemontese, M. E. 1988: *GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana.* In: *Sculoa e città* 34(3), 110–124.

McCarthy, C., Pradhan, N., Redpath, C. & Adler, A. (2016): *Validation of the Empatica E4 wristband*, 2016 IEEE EMBS International Student Conference (ISC), Ottawa, ON, Canada, 2016, 1–4.

McEwen, B. S. (2007). *Physiology and neurobiology of stress and adaptation: central role of the brain.* In *Physiol Rev*. 87(3), 873 –904.

Moneta, G. (2020). *Cognitive flow*. In Vonk, J. & Shackelford, T. (eds.), *Encyclopedia of Animal Cognition and Behavior*. Cham: Springer.

Moses, Z. B., Linda J. L, & James C. E. (2007). *Measuring Task-Related Changes in Heart Rate Variability.* In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 644–647. Lyon, France: IEEE.

Nielsen, J. (2010). What is usability? In Wilson, C. (eds.), *User experience re-mastered: Your guide to getting the right design* (3–22)*.* Burlington: Morgan Kaufmann.

Prandi, B. (2018). An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation. In Fantinuoli, C. (eds.), *Interpreting and Technology* (29–59). Berlin: Language Science Press.

Prandi, B. (2023). *Computer-Assisted Simultaneous Interpreting: A Cognitive-Experimental Study on Terminology*. Berlin: Language Science Press.

Rojo López, A. M., Foulquié Rubio, A. I., Espín López, L. & Martínez Sánchez, F. (2021) *Analysis of speech rhythm and heart rate as indicators of stress on student interpreters.* In *Perspectives 29*(4), 591–607.

Rojo López, A. M. & Korpal, P. (2020). *Through your skin to your heart and brain: A critical evaluation of physiological methods in cognitive translation and interpreting studies.* In *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 19.

Rojo López, A. M., Cifuentes-Férez, P. & Espín López, L. (2021). *The influence of time pressure on translation trainees' performance: Testing the relationship between self-esteem, salivary cortisol and subjective stress response.* In *PLOS ONE* 16(9), e0257727.

Rojo López, A. M. & Naranjo, B. (2021). *Translating in times of crisis: A study about the emotional effects of the*

COVID19 pandemic on the translation of evaluative language. In *Journal of Pragmatics* 176(April), 29–40.

Ruscio, J. (2008). *A probability-based measure of effect size: Robustness to base rates and other factors.* In *Psychological Methods* 13(1), 19–30.

Russo, M., Amato, A., Niemants, N., Torresi, I. & Spinolo, N. Forthcoming: *SmarTerp. Analysis of Students' Performance.*

Saeed, M. A., Rodríguez González, E., Korybski, T., Davitti, E. & Braun, S. (2022). Connected yet distant: an experimental study into the visual needs of the interpreter in Remote Simultaneous Interpreting. In Kurosu, M. (eds.), *HCII 2022: Human-Computer Interaction. User Experience and Behavior* (214–232). Cham: Springer Nature.

Saeed, M. A., Rodríguez González, E., Korybski, T., Davitti, E. & Braun, S. (2023). Comparing interface designs to improve RSI platforms: Insights from an experimental study. In Orasan, C., Mitkov, R., Corpas Pastor, G. & Monti, J. (eds.), *Proceedings of the International Conference HiT-IT 2023*. Naples: HiT-IT.

Schuurmans, A. A. T., de Looff, P., Nijhof, K. S., Rosada, C., Scholte, R. H. J., Popma, A. & Otten, R. (2020). *Validity of the Empatica E4 wristband to Measure Heart Rate Variability (HRV) parameters: A comparison to electrocardiography (ECG).* In: *Journal of Medical Systems* 44, 190.

Semaan, R., Nater, U. M., Heinzer, R., Haba-Rubio, J., Vlerick, P., Cambier, R. & Gomez, P. (2023). *Does workplace telepressure get under the skin? Protocol for an ambulatory assessment study on wellbeing and health-related physiological, experiential, and behavioral concomitants of workplace telepressure.* In *BMC Psychol*. 11(1), 145.

Shaffer, F. & Ginsberg, J. P. (2017). *An overview of heart rate variability metrics and norms.* In *Frontiers in Public Health* 5.

Shaffer, F., McCraty R. & Zerr, C. L. (2014). *A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability.* In *Frontiers in Psychology* 5.

Spinolo, N., Olalla-Soler, C. & Muñoz Martín, R. (2022). *Finding a way into an interpreter's heart: methodological considerations on heart-rate variability building on an exploratory study*. In *The Interpreters' Newsletter* 27, 63–87.

Stoll, C. (2009). *Jenseits simultanfähiger Terminologiesysteme: Methoden der Vorverlagerung und Fixierung von Kognition im Arbeitsablauf professioneller Konferenzdolmetscher*. Trier: WVT, Wissenschaftlicher Verlag Trier.

Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996). *Heart Rate Variability*. In *Eur Heart J* 17/28.

Jamovi project (2022). *Jamovi (Version 1.6)* [Computer Software]. https://www.jamovi.org

Veltman, H. J. A. & Gaillard, A. W. K. (1993). *Indices of mental workload in a complex task environment.* In *Neuropsychobiology* 28(1–2), 72–75.

Wadlinger, H. A. & Isaacowitz, D. M. (2011). *Fixing our focus: Training attention to regulate emotion.* In *Personality and Social Psychology Review* 15(1), 75–102.

Wang, X. & Wang, C. (2019). *Can computer-assisted interpreting tools assist interpreting?* In *Transletters. International Journal of Translation and Interpreting* 3, 109–139.

Ziegler, K. & Gigliobianco, S. (2018*). Present? Remote? Remotely Present! New technological approaches to remote simultaneous conference interpreting.*" In Fantinuoli, C. (eds.), *Interpreting and Technology*, pp. 119–139. Berlin: Language Science Press.

Appendix. Full data in aggregated form

	Session 1			Session 2		Session 3		
	Baseline	Without	With	Baseline	With	Baseline	Without	With
Mean HR	Mdn = 68.5 *SD = 15.7*	Mdn = 71.5 *SD = 18.2*	Mdn = 60.0 *SD = 9.2*	Mdn = 66.0 *SD = 7.20*	Mdn = 67.5 *SD = 14.4*	Mdn = 68.0 *SD = 10.4*	Mdn = 74.0 *SD = 13.5*	Mdn = 74.0 *SD = 10.0*
Mean RR	Mdn = 880.5 *SD = 167.5*	Mdn = 840.0 *SD = 152.9*	Mdn = 963.0 *SD =114.5*	Mdn = 910.5 *SD = 91.4*	Mdn = 890.5 *SD = 175.9*	Mdn = 844.5 *SD = 114.6*	Mdn = 801.5 *SD = 119.3*	Mdn = 812.0 *SD = 111.2*
RMSSD	Mdn = 367.3 *SD = 101.6*	Mdn = 308.3 *SD = 168.0*	Mdn = 415.5 *SD = 126.4*	Mdn = 377.2 *SD = 105.2*	Mdn = 405.1 *SD = 149.6*	Mdn = 354.6 *SD = 104.8*	Mdn = 243.9 *SD = 163.8*	Mdn = 218.5 *SD = 149.6*
LF/HF ratio	Mdn = 1.179 *SD = 0.515*	Mdn = 1.400 *SD = 0.844*	Mdn = 1.273 *SD = 0.910*	Mdn = 1.102 *SD = 0.632*	Mdn = 1.154 *SD = 0.805*	Mdn = 1.188 *SD = 1.022*	Mdn = 1.711 *SD = 2.306*	Mdn = 1.483 *SD = 1.658*