The Interpreters' Newsletter 2022 (27), 63-87
DOI: 10.13137/2421-714X/34392 ISSN 1591-4127 (print) ISSN 2421-714X (online)
https://www.openstarts.units.it/dspace/handle/10077/2119

Finding a way into an interpreter's heart: methodological considerations on heart-rate variability building on an exploratory study

Nicoletta Spinolo,
Christian Olalla-Soler and
Ricardo Muñoz Martín
DIT, Università di Bologna

Abstract

Physiological indicators of stress such as galvanic skin response, cortisol, and heart rate are gathering momentum in Cognitive Translation and Interpreting Studies. Heart-rate variability (HRV) is gaining ground as a possibly reliable indicator of stress for tasks that do not involve physical activity. However, using electrocardiography and photoplethysmography (PPG) sensors in research involves following methodological guidelines to prevent negative impacts on data. We performed an observational, exploratory study on HRV in onsite vs. remote interpreting with interpreters (n = 5) with no experience in remote interpreting. Data was collected with Empatica E4 wristbands, which use PPG sensors to measure heart rate variability. We report results, yet our focus is the methodological issues derived from using heart rate (HR) and HRV as indicators of stress that we encountered both at data collection and in the analysis. We will formulate methodological recommendations regarding HR, HRV and (1) the characteristics and size of the sample; (2) the structuring of data collection sessions; (3) the selection of stimuli; (4) its relationship with other variables; (5) the selection of heart-related indicators; and (6) statistical analysis.

Keywords

Interpreting, arousal, heart-rate variability, methodology, naturalistic data collection.

Introduction and rationale

Research on remote and distance interpreting has compared stress levels of professionals carrying out onsite and distance simultaneous interpreting with a traditional interpreting booth and console (Moser-Mercer 2005, Roziner/Shlesinger 2010). Here, we compare stress levels of professionals carrying out onsite vs. remote simultaneous interpreting (RSI) tasks on an RSI platform. Stress was measured with the indicators heart rate (HR) and heart-rate variability (HRV, time changes between consecutive heartbeats). The research questions we explored are:

1. Is there a difference between remote SI and onsite SI in terms of stress?
2. If so, does remote SI become less stressful after the first use of an RSI platform?
3. Does stress correlate with (a) filled pauses in the target text, (b) other disfluencies in the target text, and (c) delivery speed of the source text?

In this exploratory study, our aim was to test our research design (§2) and not to provide answers to these questions. Hence, we provide our results (§3), discuss the methodological limitations, and provide recommendations for using HR and HRV as indicators of stress (§4).

1. Stress

Professional *multilectal mediated communication* (MMC) tasks–especially, simultaneous interpreting (SI)–are customarily described as *stressful* activities (e.g., Mackintosh 2003; Bayer-Hohenwarter 2009; Gile 2009: 112–113) because they often come with time pressure and high cognitive demands. In the Cognitive Translation and Interpreting Studies (CTIS) literature, the terms *stress, time pressure* and *anxiety* are sometimes used interchangeably but some terminological precisions are in order.[1] These terms pivot on *arousal*–i.e., how awake, activated and alert you are at a point in time; hence, *arousal* refers to graded states of neurovegetative activation associated with physical and psychological variations. Arousal is a coin with a physical side and a psychological side. Physically, arousal involves biological systems such as the autonomic nervous system and the endocrine system. Psychologically, arousal impacts emotions and behavior, and it affects memory, attention, and decision-making. The intensity of arousal may span from *apathy* to *relaxation*, *stress*, and *anxiety* (Figure 1). Non-pathological *apathy* describes a lack of motivation or goal-directed behavior and indifference to one's surroundings, and *relaxation* hints at abatement of intensity, vigor, energy, or tension, resulting in calmness of mind, body, or both (VandenBos 2015).

1 *Time pressure*–the feeling or awareness that one's duties exceed one's ability to complete them in the afforded time–is a well-known kind of stress for the MMC professional (e.g., Jensen 1999; De Rooze 2003; Sharmin *et al.* 2008; Weng/Zheng 2020; Rojo *et al.* 2021; Weng *et al.* 2022). It should not be confused with *lack of time*, which is the stressor prompting time pressure (see below).

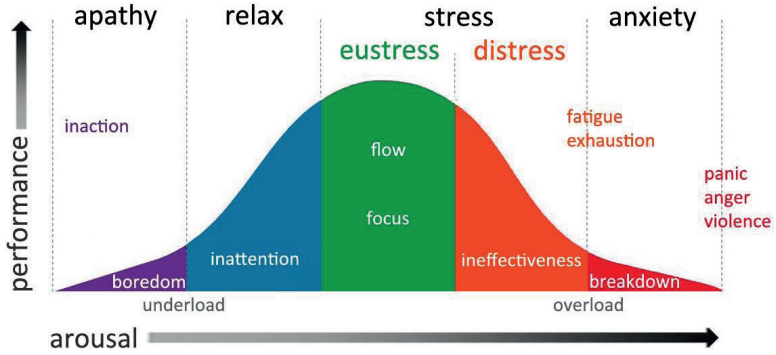

Figure 1. A model of the Yerkes-Dodson (1908) law of interaction between arousal and performance

Stress is a normal response of emotional tension to *stressors* (see below). Moderate stress, or *eustress,* is positive, whether in bursts–as when reacting to avoid a sudden danger–or continuous, as when we work feeling focused and in control. Eustress may foster a state of deep concentration and task absorption known as *cognitive flow* (Moneta 2018), which often results in both an optimal experience and top performance (Stranks 2005: 5). When facing difficulties, higher levels of stress may result in feeling overwhelmed or unable to cope with standing demands. This is *distress,* a non-adaptive response that may result in impaired performance, blocks and breakdowns.[2] *Anxiety* is a strong reaction to distress, a future-oriented, long-acting emotion of apprehension, fear, dread, or uneasiness and somatic symptoms of tension that anticipate impending danger or misfortune (cf. VandenBos 2015). Anxiety may linger after its cause wanes or disappears.

Stressors are detrimental or damaging factors causing stress. Cooper *et al.* (1982) class interpreters' stressors into physical/environmental, task-related, and interpersonal. Examples of stressors are poor ventilation, lighting problems, booth space, acoustics, sustained concentration, delegates reading their speeches, having a heavy accent or speaking too low, and the like. Stressors researched in CTIS include long turns in the booth (Moser-Mercer *et al.* 1998), conditions in media interpreting (Kurz 2002), lack of background knowledge (Macintosh 2003), and fast speech (Korpal 2016).

Regardless of the stressor, the nature (not the intensity) of individual response patterns tends to remain unaltered. However, people react in different ways when facing a stressor. De Rooze (2003) found systematically lower translation quality in most translators under stringent time constraints, but 25% of them reached higher quality scores also systematically (three data-collection in-

2 As in Figure 1, we keep the terms *eustress* and *distress* to underscore that the response may be deemed positive or negative, but stress seems to be just one phenomenon, albeit graded (Bienertova-Vasku *et al.* 2020), and part of the wider construct of arousal. Hence, it would be more correct to describe eustress as *positive arousal* and distress as *negative arousal.* Both positive and negative arousal may be moderate or high.

stances).[3] People perceive and adapt to stressors in different ways (Sapolsky 2015; Ebner/Singewald 2017).

Stressors may be situational, like noise and the place where you interpret. Moser-Mercer (2005) studied interpreters' physiological reactions in on-site and RSI and found that RSI was more stressful, and performance quality decayed faster. Roziner/Shlesinger (2010) also found a faster decline in quality but no significant difference between both interpreting situations, except as self-reported. Nevertheless, it is unclear whether such differences are due to the task or to the setting. Cooper *et al.* (1982) also mention that stressors vary between home and work environments, such that one scenario may sooth the stress away from the other one or spill over and add to it. Onsite and RSI differences merit more study, in view of the recent massive move towards remote communication.

Stress is technically a physical response to a disruption of *homeostasis* (Billman 2020)–the dynamic, adaptive state of good physical and chemical balance in physiological processes and bodily functions (e.g., body temperature, blood pressure, blood sugar and pH levels, etc.). Homeostasis is achieved through self-regulating mechanisms that overcome a natural resistance to change, to adjust to personal and environmental demands. Emotional responses to stress include anxiety, restlessness, frustration, helplessness, fear, anger, sadness, and disgust. Questionnaires are often used for exposure to stressors over time (e.g., adult STRAIN, Slavich/Shields 2018) and momentary emotional responses (e.g., STAI Y1, Spielberger *et al.* 1983; PANAS-SF, Thompson 2007), and they have been used in CTIS (e.g., Courtney/Phelan 2019, Korpal 2021). However, correlations between self-reported measures and physiological indicators are poor (Hellhammer *et al.* 2010: 189; cf. also Moser-Mercer 2003).

Physiological indicators measure physical arousal. In CTIS they include blood test for immunoglobulin M (Moser-Mercer 2003), salivary cortisol test (Moser-Mercer *et al.* 1998; Mackintosh 2003; Moser-Mercer 2003; Roziner/Shlesinger 2010; Rojo *et al.* 2021), blood pressure (Klonowicz 1994; Mackintosh 2003; Roziner/Shlesinger 2010; Korpal 2016; Baghi/Khoshsaligheh 2019), heart rate (e.g., Klonowicz 1994; Kurz 2002, 2003; Mackintosh 2003; Roziner/Shlesinger 2010; Korpal 2016; Baghi/ Khoshsaligheh 2019; Rojo *et al.* 2021), and electrodermal activity (Matamala *et al.* 2020). Taking blood samples is very invasive and may deter participation, and cortisol takes 20 to 30 minutes to peak after momentary or acute stressor onset, and about 60 to return to normal. Thus, cortisol seems adequate to scrutinize the overall stress experienced in a whole task. Blood pressure also makes it difficult to determine punctual sources of stress (Korpal 2016: 311). Korpal (2016) suggests that blood pressure works better as a trait indicator but not for momentary, situation-induced stress (see also Hjortskov *et al.* 2004). Gordon/Berry (2021) use new indicators (blood pressure changes and reactivity) that hold promise.

Stress stimulates the heart both via the sympathetic and parasympathetic autonomic nervous systems. The sympathetic activity raises our heart rate (e.g.,

3 No data was collected as to how participants experienced the challenge, but eustress and distress might well be the reason for the difference. As for cognitive flow, it might explain instances of translation *peak performance* (cf. Jakobsen 2005).

when doing physical exercise), while parasympathetic activity lowers it (e.g., when resting). HRV reflects the dynamics of such interaction and provides a measure for stress through the activity of the autonomic nervous system. Low parasympathetic activity, characterized by a decrease in the high-frequency band and an increase in the low-frequency band, is the most frequently reported factor (Kim *et al*. 2018). There is some progress in determining statistical correlations between the average HRV and the intensity of stress indicators and best heartbeat-related indicators in short term mental stress (Fauquet-Alekhine *et al*. 2016; Hu/Gao 2022). Crucially, wearable sensors (especially, wrist-worn devices) can measure daily life stress (Kyriakou *et al*. 2019) and are used in CTIS (Weng/Zheng 2020; Weng *et al*. 2022). The unrestricted movements of participants wearing these devices can introduce artifacts that can be filtered out (Can *et al*. 2019).

2. Materials and methods

The dependent variables in our exploratory, intra-subject quasi-experimental study were (a) stress, measured with HR and HRV (see §2.4.1), and (b) performance indicators (filled pauses and other disfluencies–truncated words, reformulations, repetitions, false starts; §2.4.2). Our independent variables were (I) the interpreting setting (onsite vs. remote) and (II) exposure (first vs. second use of RSI platforms).

We first describe the procedure (§2.1) and our sample (§2.2). Then we describe our data-collection tools, Empatica E4 wristbands (§2.3.1) and audio recordings (§2.3.2). Next, we overview HR and HRV indicators (§2.4.1) and performance (§2.4.2), and data-extraction procedures and we describe the source speeches (§2.5). The statistical procedures we used are in §2.6, and we conclude the section with remarks on ethics (§2.7).

2.1 Procedure

We conducted the study in February 2020. The convenience sample of professionals convened twice for sessions held two weeks apart at university premises. In each session, they interpreted two live speeches (ES > IT) in different settings. For each interpretation, a moderator introduced the Spanish L1 speaker (always the same one).

The first session–*onsite interpreting* setting–was held at a lecture hall with booths with traditional *hard consoles* and direct view of both speaker and moderator. The second session–*remote interpreting* setting–was held at an interpreting lab; participants interpreted in booths with computers connected to the RSI platform Voiceboxer.[4] The moderator and the speaker connected to it from another booth, but interpreters could only see the speaker on screen, through the RSI platform. The order of the two settings (onsite first and remote second) was the

4 <https://voiceboxer.com/>.

same in both sessions; in a larger study with a larger sample, the order of conditions should be counterbalanced to control for fatigue effects. In this case, we knew that our sample size would not allow any conclusions, so we opted for practicality in data collection and on piloting a methodology for future larger studies.

During the sessions, this was the procedure:

1. Participants welcomed to lecture hall.
2. [first session] Participants orally briefed about the study, with the chance to ask questions. Informed consent forms signed.
3. Empatica E4 wristbands placed on participants' wrists and turned on. The researchers noted down times when the wristbands were turned on.
4. Participants proceeded to the booths, and were given some time to get used to the room and adjust the settings of the hard console.
5. The moderator introduced the speaker and the topic, then the speaker read the speech. The speech exact starting and ending times were noted down.[5]
6. Once finished, participants had some time to relax. Then they were given a source-text transcript to mark all segments that had drawn their attention for some reason (difficulty, problems, etc.).
7. Participants were walked to the remote interpreting setting. Once there, they were placed in the booths and given 15-20 minutes to relax. This timespan was used to measure the participants' HR and HRV baseline (see §2.4.1).
8. In the first session with the RSI platform, participants were introduced to platform functions and controls.
9. Step 5 was repeated.
10. Then step 6 was repeated. Relax time was used to measure the participants' recovery phase (see §2.4.1), after completion of both tasks.

After step 10, the session ended. In all, each session lasted 60–90 minutes. Data collection for both sessions was in two turns, due to the diverse availability of participants. Having the speech delivered live increased ecological validity, even though it was not possible to control for delivery speed in different turns. However, the mean and median speed duration was homogeneous in all turns (see §2.5).

2.2 Sample

Using a non-probabilistic, convenience sampling procedure (personal contacts), we recruited seven participants (five females, two males) with >5 years of professional experience but none with RSI platforms.[6] Demographic data on partic-

5 The Empatica E4 wristbands include an 'event marker' function, activated through a button that may also turn off or reset the device if pressed too long. To prevent data loss, we chose to note down the timings and ask participants not to touch the wristband. This also contributed, we believe, to make the participants 'less aware' of the wristband.

6 As of April 2022, this condition is difficult to meet in Italy, but it was not so in February 2020, right before the start of the pandemic.

ipants were collected in advance. Data of two participants from the first session was lost due to technical problems. Due to within-subject design, all their data was discarded. The remaining five participants (one male, four females) had a mean age of 39 (Mdn = 36, min = 34, max = 47), Italian as A language and Spanish and English (and one of them Portuguese) as B languages. Their mean professional experience was 12.8 in years (Mdn = 10, min = 5, max = 23) and 2760 in hours (Mdn = 3000, min = 1000, max = 4800).

2.3 Data-collection tools

2.3.1 Empatica E4

Empatica E4 wristbands are equipped with sensors that monitor physiological reactions to stimuli:

- A photoplethysmography sensor measures HR, blood volume pulse (BVP), and inter-beat interval data.
- A galvanic skin response sensor registers electrodermal activity.
- An infrared thermopile measures peripheral skin temperature.

We used BVP data to extract HR and HRV indicators (see §2.4.1).

2.3.2 Audio recordings

Interpretations and source speeches were audio-recorded in both sessions. Given the chosen indicators (see §2.4.2), data were directly extracted from the files, rather than from their transcripts.

2.4 Indicators

2.4.1 Heart rate and heart-rate variability

As explained, we decided to focus on HR and HRV as indicators of stress. HRV is not a single measurement, but can be assessed through several measurements. Castaldo *et al.*'s (2015) meta-analysis focused on short-term HRV recordings to identify the most reliable indicators of mental stress and conclude that "the pooled values of 7 HRV measures (RR, SDRR, RMSSD, pNN50, D2, HF and LF/HF) out of the 9 meta-analyzed changed significantly during mental stress" and "should be considered as possible pivot values for future studies" (2015: 376).

indicator	definition[a]	significant trend under stress[b]
mean RR	Mean duration of RR intervals (where R is a peak of the QRS complex of the ECG wave)	lower
SDRR	Standard deviation in time between RR intervals	lower
RMSSD	Root mean square of successive differences between normal heartbeats	lower
pNN50	Percentage of adjacent NN intervals that differ from each other by more than 50 ms (where NN intervals are intervals between normal R peaks)	lower
D^2	Correlation dimension	lower
HF	High frequency power	lower
LF/HF	Ratio between LF and HF band powers	higher

[a] Based on Schaffer *et al.* (2014) and Schaffer/Ginsberg (2017).
[b] After Castaldo *et al.* (2015).
Table 1. HRV indicators of mental stress, definitions, and their behavior

Following Rojo *et al.* (2021), we analyzed HR and several measures for HRV, as Castaldo *et al.* (2015) suggested. Table 1 lists these HRV measures, what they indicate and how they tend to vary during mental stress. HRV measurements were extracted with the Kubios HRV software, following the criteria of the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (henceforth, Task Force 1996).[7] SDRR– not extracted by Kubios–was not included.

Following Rojo/Korpal (2020), we extracted 5-minute spans from each condition, i.e., baseline (resting, no task), onsite interpreting (task 1), RSI (task 2), and recovery (after tasks), following the rule of the three Rs (Resting-Reactivity-Recovery; Laborde *et al.* 2017). To minimize confounders due to stress induced by starting the experiment and by fatigue at the end of the task, we took the 5 central minutes of each condition (Task Force 1996). Kubios HRV was used for automatic artifact detection and correction (Lipponen/Tarvainen 2019). A mean of 10.3% (Mdn = 9.9; SD = 1.2) of the beats were corrected in each recording.[8]

7 <https://www.kubios.com/>.
8 Percentages of non-corrected beats: subject 1, session 1: 91.4%; subject 1, session 2: 90.1%; subject 2, session 1: 90.3%; subject 2, session 2: 88.4%; subject 3, session 1: 89.1%; subject 3, session 2: 88.3%; subject 4, session 1: 91%; subject 4, session 2: 90%; subject 5, session 1: 88.9%; subject 5, session 2: 90.1%.

2.4.2 Performance indicators

Performance analysis focused on filled pauses and other disfluencies, i.e., "phenomena that interrupt the flow of speech and do not add propositional content to an utterance" (Gósy 2007: 93). Here, other disfluencies were truncated words, reformulations, word repetitions, and false starts, counted as a single category *(other disfluencies)* by two raters. In interpreting, fluency (and lack thereof) has been linked to quality (Pradas 2006) and to cognitive load (Tóth, 2011; Plevoets/Defranq, 2016; Defranq/Plevoets, 2017; Bóna/Bakti 2020).

Source speeches had not been pre-recorded, so their speed could differ between sessions, but almost no variation was detected (see §2.5). The number of filled pauses and other disfluencies varies depending on the length of the speech: the longer the text, the more likely it is that the number of filled pauses and other disfluencies will increase. These two indicators are affected by the speed of the input speech. Consequently, we normalized both indicators: counts of filled pauses and of other disfluencies were divided by the source text speed (words per minute).

2.5 Source texts as stimuli

One speaker delivered all four speeches live in Spanish, and participants interpreted them into Italian. The speeches, prepared specifically for the study out of popular science websites and blogs, covered similar topics: (I) climate change and migration; (II) climate change, migration, and gender; (III) EU policies on climate change; and (IV) climatic migrations. Text order was randomized for data collection turns. However, the sequence of conditions (onsite and online) was not randomized. The mean duration of the speeches was 12.6 minutes (median 12.5), mean word count was 1514 (median 1503) and mean speed was 121 wpm (median 120). Original speech scripts were similar in word-count (1440, 1519, 1623, and 1502) but they were read live, so it was not possible to control for speed and duration. Nevertheless, the mean and the median are almost equal in both speech duration and mean speed, so these two factors vary very little, which enhances the comparability of all four conditions.[9]

2.6 Statistical analysis

Laborde *et al.* (2017: 12) recommend to log transform the collected data, given that HRV parameters frequently present a non-normal distribution. Log transformations may reduce the skewness of data, bringing them closer to normal distributions which allow for the use of parametric statistics (but see Feng *et al.* 2014 for arguments against the use of this procedure). We only had 10 datapoints to compare onsite vs. remote interpreting and just 5 to compare first vs. second

9 The four speeches can be accessed here: <https://doi.org/10.5281/zenodo.7101770>.

session for onsite and remote interpreting. Hence, log transforming data would not significantly improve the statistical quality. With such a small dataset, even if log transformed, parametric procedures are not adequate. Consequently, we applied three non-parametric tests:

- Wilcoxon Signed-Rank test, to compare two conditions on one indicator (i.e., onsite vs. remote or 1st session vs. 2nd session).
- Friedman's test, to detect differences between the four measurements (baseline, onsite, remote, and recovery) for a given indicator in a given session.
- Kendall's Tau B, to detect correlations between HR and HRV indicators and performance indicators.

Statistical analyses were performed with Jamovi 2.3.0 with a pre-established significance level of 0.05.[10]

2.7 Ethical issues

The informed consent form explained that the study targeted interpreters' reaction while using different working tools. It included information on the nature of collected data, interpreters' performances and Empatica E4 data on cardiac and electrodermal activity. It also informed participants that anonymized data would be always presented in aggregate ways. Participants were informed that they could withdraw from the study at any moment and their data would be deleted with no consequences for them. All participants signed the informed consent form. They received financial compensation for participating in the study.

3. Results

3.1 Differences in stress between onsite and remote interpreting

Descriptive statistics show very similar mean and median values for all indicators (Table 2). The largest difference is in RMSSD, which is lower (i.e., more stress) in remote than in onsite interpreting.[11] This is the only indicator with some variation between the two conditions. No indicator presented a statistically significant difference, i.e., no differences were detected between onsite and remote interpreting.

10 <https://www.jamovi.org/>.

11 The difference in D^2 also seems relevant when comparing the means, but not the medians.

indicator	type	Mean (m)	Median (Mdn)	SD	Wilcoxon Signed-Rank test		
					Z	*p*	*r*
Mean HR	onsite	68.80	66.50	9.67	31.0	0.759	0.127
	remote	68.10	67.00	5.71			
Mean RR	onsite	885.80	899.50	113.10	26.0	0.919	0.055
	remote	886.50	894.50	70.91			
RMSSD	onsite	354.88	396.80	135.76	34.0	0.557	0.236
	remote	330.97	378.95	134.67			
pNN50	onsite	70.61	80.52	25.67	23.0	0.695	0.164
	remote	70.86	80.47	20.44			
D^2	onsite	0.63	0.00	1.32	2.0	0.093	0.810
	remote	1.07	0.01	1.78			
HF	onsite	45.67	48.58	13.56	23.0	0.999	0.022
	remote	46.35	46.51	14.72			
LF/HF	onsite	1.56	1.06	1.50	23.0	0.999	0.022
	remote	1.40	1.15	0.90			

Note: The tests were bilateral.
Table 2. HR and HRV indicators by interpreting mode and comparison between the two modes (N=10)

We found no difference in disfluencies and filled pauses either between conditions (Table 3), thus we did not detect any differences in stress between onsite vs. remote interpreting.

Normalized…	Onsite			Remote			Wilcoxon Signed-Rank test		
	m	Mdn	SD	M	Mdn	SD	*Z*	*p*	*r*
pauses	0.30	0.33	0.14	0.27	0.23	0.17	36.0	0.432	0.309
disfluencies	0.26	0.20	0.13	0.24	0.23	0.08	36.0	0.432	0.309

Note: Tests were bilateral.
Table 3. Normalized pauses and disfluencies in onsite and remote interpreting and comparison between the two modes (N=10)

3.2 Differences between first and second session

Descriptive statistics for HR and HRV indicators yield very little variation between sessions (Table 4) and the Wilcoxon Signed-Rank test confirmed non-significant results. Hence, there seems to be no decrease in stress in the second session, whether onsite or RSI. No differences were detected between sessions in

pauses and other disfluencies in both conditions (Table 5); we therefore found no decrease in stress in the second session.

						Wilcoxon Signed-Rank test		
indicator		session	m	Mdn	SD	*Z*	*p*	*r*
Mean HR	onsite	1	69.00	64.00	11.79	7.5	0.554	0.001
		2	68.60	69.00	8.44			
	remote	1	67.40	66.00	5.32	6.0	0.699	0.200
		2	68.80	68.00	6.61			
Mean RR	onsite	1	888.20	932.00	132.15	7.0	0.594	0.067
		2	883.40	867.00	106.31			
	remote	1	894.00	913.00	70.91	9.0	0.406	0.200
		2	879.00	886.00	78.39			
RMSSD	onsite	1	342.04	405.70	176.96	7.0	0.594	0.067
		2	367.72	358.10	98.71			
	remote	1	335.40	377.50	161.25	7.0	0.594	0.067
		2	326.54	380.40	121.47			
pNN50	onsite	1	68.69	81.21	34.20	8.0	0.500	0.067
		2	72.53	73.11	17.44			
	remote	1	71.90	82.83	24.46	7.0	0.594	0.067
		2	69.81	75.31	18.41			
D^2	onsite	1	0.51	0.00	1.15	0.0	0.969	0.000
		2	0.74	0.01	1.61			
	remote	1	0.93	0.00	2.08	3.0	0.819	0.400
		2	1.21	0.27	1.66			
HF	onsite	1	47.78	55.67	19.26	10.0	0.313	0.333
		2	43.56	42.19	5.65			
	remote	1	49.88	48.05	14.32	11.0	0.219	0.466
		2	42.83	41.33	15.85			
LF/HF	onsite	1	1.79	0.80	2.21	5.0	0.781	0.333
		2	1.32	1.37	0.30			
	remote	1	1.17	1.08	0.75	2.0	0.938	0.733
		2	1.64	1.42	1.07			

Note: T tests were unilateral.
Table 4. Indicators in onsite and remote interpreting by session and comparison between the two sessions (N=5)

		session 1 (N = 5)			session 2 (N = 5)			Wilcoxon Signed-Rank test		
		m	Mdn	SD	M	Mdn	SD	*Z*	*p*	*r*
pauses	onsite	0.29	0.31	0.14	0.31	0.36	0.16	6.0	0.688	0.200
	remote	0.24	0.23	0.09	0.30	0.22	0.22	6.0	0.688	0.200
disfluencies	onsite	0.22	0.17	0.10	0.30	0.24	0.16	2.0	0.938	0.733
	remote	0.21	0.22	0.08	0.26	0.23	0.08	3.0	0.906	0.600

Note: Tests were unilateral.
Table 5. Normalized pauses and disfluencies in onsite and remote interpreting by session and comparison between the two sessions

3.3 Correlation between stress and (a) target text filled pauses, (b) other target text disfluencies, and (c) source text delivery speed

Small, statistically significant correlations were detected between filled pauses and RMSSD, pNN50, and D^2 (Table 6):

- The larger the RMSSD (less stress), the more filled pauses.
- The larger the pNN50 (less stress), the more filled pauses.
- The lower the D^2 (more stress), the more filled pauses.

These results are not conclusive. First, only three of the HR and HRV indicators yielded statistically significant correlations. Second, the three indicators with such correlations showed small strengths of association. Third, in RMSSD and pNN50, the positive association between indicators seemed counterintuitive, since more filled pauses might be expected with higher stress. Fourth, filled pauses is the only performance indicator showing some association with stress-related indicators.

Kendall's Tau B	normalized pauses	normalized disfluencies	WPM
Mean HR	-0.321	-0.043	-0.078
Mean RR	0.295	0.032	0.077
RMSSD	0.358* (p = 0.028)	0.095	0.109
pNN50	0.358* (p = 0.028)	-0.011	0.131
D^2	-0.343* (p = 0.047)	0.042	-0.275
HF	0.121	0.069	0.137
LF/HF	-0.121	-0.069	-0.137
WPM	0.077	-0.164	–
Normalized disfluencies	0.211	–	–

Table 6. Correlations between HR and HRV indicators and filled pauses, other disfluencies, and delivery speed

Consequently, we cannot answer our research question. Our results suggest that the root causes of stress and their manifestations are difficult to identify and interpret, so the relationship between stress and performance indicators deserves further study.

4. Discussion and methodological suggestions

As mentioned, this study aimed at testing our design and refining it for future studies using HR and HRV indicators. Below we present methodological recommendations based on the main limitations and constraints.

4.1 Sample size and statistical power

Studies on physiological stress measured with HR and HRV indicators are intra-subject by default, even if the design can be expanded with inter-group comparisons (see §4.2). In such studies, a single datapoint is computed for each indicator, measurement, and participant, so small sample sizes cannot be compensated with multiple datapoints. Hence, small samples will render designs underpowered even though intra-subject designs increase statistical power (Lakens 2022). This is our case, where a sample of five participants was not powerful enough to detect significant differences between tasks and sessions.

Recommending samples as large as possible is nothing new. The difficulties in finding professionals willing to contribute to such studies are not new either, especially if they need to commute to an academic institution to participate, when data collection takes place over several sessions, or when funds are not available to compensate them financially. If sample sizes have not increased after years of calls to do so, it may not be due to a lack of willingness on the part of researchers. Our recommendation still is to achieve as large a sample size as possible to ensure sufficient statistical power to detect meaningful differences between the conditions being compared.[12] Sample sizes can be determined by establishing a desired (1) level of statistical power (i.e., the likelihood of rejecting the null hypothesis when it is false; also called β or Type II error); (2) significance level (the likelihood of rejecting the null hypothesis when it is true; also called α or Type I error), and (3) minimum effect size of interest.

The scarce CTIS studies using HR and HRV indicators make it difficult to conclude what the minimum effect size of interest might be. This scarcity also makes

12 The meaningfulness (i.e., relevance) of a relation between two variables or between (at least) two groups is measured through effect sizes. Statistically significant results with small effect sizes may render the difference or correlation meaningless. This is why in recent years there have been calls for reporting effect sizes and interpreting them together with statistical significance (i.e., *p*-values) when drawing conclusions from statistical analyses (Hedges 2008; Rosnow/Rosenthal 2009; Mellinger/Hanson 2017). Pre-establishing a level of meaningfulness means that, below that pre-established level, a result will not be considered relevant even if it is statistically significant.

meta-analyses impossible, so researchers have two main options. The first one is to use the general guidelines and thresholds to interpret effect sizes (see Cohen 1988). Such thresholds are arbitrary and may not be appropriate in all contexts (e.g., when a meta-analysis has provided a minimum effect size of interest which differs from the threshold for small effect sizes).

The second option is to search the literature for studies using the same HR and HRV indicators in comparable situations of stress.[13] Meta-analyses are especially useful. Quintana (2017) carried out a meta-analysis with 297 HRV effect sizes and concluded that effect sizes of $d = 0.25$, $d = 0.5$, and $d = 0.9$ should be interpreted as small, medium, and large. Hence, depending on the aims of their study (i.e., whether exploratory or confirmatory), their limitations–i.e., financial, time-related, population-related, etc.–and the level of precision needed, researchers may decide to look for a minimum effect size of interest of 0.25, 0.5, or 0.9. Free software packages such as G*Power 3.1 (Faul *et al*. 2009) will automatically compute the minimum sample size needed to detect a given minimum effect size of interest at specific significance and power levels (which are generally set to 0.05 and 0.80). Yet such heuristics have been contested and Maier/Lakens (2021) call to adapt and justify such levels to the characteristics of each study.

4.2 Expertise and stress endurance

As expertise increases, participants are likely to be more used to endure stress and adapt quickly to new situations. Moser-Mercer (2005) and Roziner/Shlesinger (2010) studied experts and found no significant differences in physiological stress between onsite and RSI, although self-reported measurements did show more stress in RSI. Both studies found an earlier onset of fatigue in RSI, indicated by a faster deterioration of performance.

Consequently, samples with different levels of expertise would have been necessary in order to control expertise as an independent variable when comparing the participants' physiological stress (e.g., comparing interpreting trainees or recent graduates with more experienced interpreters). However, studies aiming at replicating this design will find it even harder to recruit participants with no experience on RSI. Due to the pandemic, most students and professional interpreters may now be even more used to RSI than to onsite interpreting. For research questions other than those contrasting onsite vs. RSI settings, studies on physiological stress should isolate expertise as an independent variable, so that differences (or a lack thereof) can be better addressed.

13 For instance, effect sizes reported in HR and HRV studies on sports science may not be adequate for a study on mental stress.

4.3 Structure and duration of data collection sessions

Data collection sessions for HR and HRV measures of stress should follow the guidelines of the Task Force (1996). Each task needs to be long enough to extract a 5-minute-long central span. In order to extract data of phases of the same length, researchers need to design the sessions so as to record baseline and recovery phases of at least five minutes each.

With at least three phases per session (baseline, stimulus, recovery), data collection sessions can become tiring for participants, especially when exposed to demanding stimuli. Researchers planning to include more than two stimuli (e.g., speeches) should consider adding breaks between them, so that other uncontrolled variables such as fatigue have the smallest effect possible. Randomizing the order of the stimuli will also help reducing undesired effects of confounders in long data collection sessions.

4.4 Optimal HRV measurement

Multiple factors impact HR, such as gender, body mass, age, health conditions, food and caffeine intake, and body movement (comprehensive list in Laborde *et al*. 2017: 6–7). To obtain clean data, researchers should control for as many factors as possible. Some of them can be controlled by screening participants, to build a sample with similar characteristics (e.g., selecting participants with ages within a pre-established age range), and other factors–such as not having meals and caffeine beverages in the two hours before the session–need to be presented as instructions to participants (Laborde *et al*. 2017: 7). Some factors can be used as covariates in regression analyses.

The stress levels interpreters feel at work are not constant. There may be peaks at specific segments of the speech, due to several reasons, like a sudden increase of delivery speed, terminologically dense passages which require documentation, complex grammatical and syntactical structures, etc. Stress may be higher at the beginning of the speech, when the interpreter is getting familiar with the topic and the speaker's accent and speed. Given this fluctuation, researchers need to create their stimuli such that, during the five central minutes of the task, the input will induce a (somewhat) constant state of stress.

Controlling for movement or asking participants to sit still may be counterproductive, and it drastically lowers ecological validity (see §4.5). Here we focus on the optimal collection of HRV data in the three measurement types: baseline, task, and recovery phase.

To measure the baseline, Laborde *et al*. (2017: 9) recommend that the participant be seated with both feet on the floor, ankles at an angle of 90°, hands on their thighs, and eyes closed. Other positions are possible, as long as it is the same position during exposure to the stimuli. To get used to the setting, the participants should be in this position for about five minutes. The recovery phase begins right after the end of the exposure to stimuli, and the position of the participants should again be as close as possible to the one in the baseline.

Baselines are generally measured before the study task, but before interpreting, participants may feel stress or anxiety due to task pressure. Hence, measuring the baseline before the task might not provide an accurate measurement of the participants' resting state, and differences between baseline and on-task measurements may be small. In studies on mental stress, the effects of pre-task stress on baseline measurements should be further investigated and compared to post-task baseline measurements. Post-task baselines could be a more accurate indicator of the participants' resting state. Moses *et al.* (2007) conducted a study on mental stress and observed that HRV measurements returned to baseline levels within a five-minute recovery period. Conversely, pre-task measurements of the baseline may be more adequate in studies involving a change in the participants' physical state before and during the task due to physical exertion (e.g., sports physiology). This empirical hypothesis would require further investigation to determine how to acquire accurate baseline measurements in cognitively demanding tasks not involving physical exertion.

4.5 Ecological validity

Researchers need to find a tradeoff between ecological validity and control. Here, guaranteeing a correct use of the wristbands and the adequate unfolding of the data collection procedures required that the participants came to our premises. This reduced ecological validity, especially in the RSI setting: remoteness was simulated but participants were aware that they were all in the same lab at the same time.

Another limitation is the lack of an audience: the lecture hall for the onside condition was almost empty, excluding the speaker, the moderator, a member of the research team and a technician, and no audience was connected to the remote session. This might well have influenced the participants' perception of the event and very probably their stress reaction, since the experimental setting was obvious.

As explained in §4.4, when collecting data through a PPG sensor, data can significantly deteriorate due to participants' movements introducing artifacts. Simultaneous interpreting does not require much movement, but interpreters do move in the booth to use the console, take notes, etc. Stress and concentration can also prompt movements and gestures. Participants should wear the wristband on their non-dominant hand, which is likely to move less during the task, especially if it involves actions like note taking–but this will not completely eradicate the problem of artifacts.

Finally, participants had no boothmate in any setting. While the speeches were short enough to be interpreted by just one participant, boothmate interaction is an important component of simultaneous interpreting, and one that differs very much between the onsite and RSI settings, especially when boothmates are not co-located.

4.6 Data Triangulation

As explained in §3, the mean HR and HRV indicators yielded no significant differences in stress. Furthermore, HR and HRV data were cross-referenced with performance indicators, but they should also be cross-referenced with self-reported measurements, not only to spot differences between the tasks, but also to understand how participants cope with stress. For instance, an increase in physiological stress in SI could link to increased attention or monitoring one's own performance in order to maintain quality. Self-reported data can also be useful to understand to what extent the participants' perception of stress correlates with the physiological measures. Psychometric questionnaires, such as the STAI (Spielberger *et al*. 1983; see, Rojo *et al*. 2021), or post-task interviews can be used to gather this type of data.

4.7 Analyzing, interpreting, and reporting HRV data

HR and HRV data are generally non-normally distributed and we also recommend to log transform them (Laborde *et al*. 2017: 12). Given the intra-subject design of HR and HRV studies, tests for related samples are necessary (Mellinger & Hanson 2017 review tests comparing two and more related samples). Laborde *et al*. (2017: 12) also recommend using autoregressive models to analyze frequency-domain indicators. Quintana *et al*. (2016) developed a set of guidelines for the adequate reporting of HRV research in psychiatry. We have adapted their guidelines for CTIS research.

Topic	Checklist item
Minimum effect size of interest	Provide information about the procedure followed to identify the minimum effect size of interest (e.g., in meta-analyses, previous studies, etc.). Justify the minimum effect size of interest in relation to the characteristics of the study (e.g., the aims, the design type, etc.).
Sample size determination	Discuss the constraints faced by the researchers (e.g., lack of funding, small population, lack of personal resources, etc.) that may affect the size of the sample. Report the significance level. Report the power level. Report the results of the sample size calculation. Report any adjustments to the results of the calculation.

Table 7a. Guidelines for reporting HRV research in CTIS: *Sample size*

Topic	Checklist item
Group selection criteria	List the inclusion criteria (e.g., years of experience) and justify them when necessary. Provide information about the way the inclusion criteria have been checked (e.g., through a selection questionnaire, an interview, etc.).
Demographics	Report details of the academic and/or professional characteristics of the group(s), such as year of study, years of experience, language combinations, etc. Report details of factors that may affect HR and HRV indicators (such as age, gender distribution, physical activity level, nicotine and alcohol intake, etc.).

Table 7b. Guidelines for reporting HRV research in CTIS: *Selection of participants*

Topic	Checklist item
Research questions / hypotheses	List the research questions or hypotheses.
Type of design	Present the type of design (experimental, quasi-experimental, observational, etc.; inter- or intra-subject). Discuss the adequateness of the design to answer the research questions or hypotheses.
Structure of the sessions	Describe the data-collection sessions and the measures. Describe the setting where the study took place.
Texts used as stimuli	Describe thoroughly the texts used as stimuli (length, text profiling indicators, manipulations, etc.). If the exposure to the stimuli has been randomized, describe the procedure.
Self-reported measures (if any)	If the HR and HRV indicators are complemented with self-reported measures, present and describe them. If using an existing psychometric test, justify the selection. In the case of creating a new test, report its psychometric properties.
HR and HRV collection tools	Describe the tools used to collect HR and HRV data. If using electrodes, describe the electrode configuration.
HR and HRV collection details	Report the length of the recordings. Provide information about the way the baseline was measured (participants' position, duration, previous acclimatization, etc.).
Selection of HR and HRV indicators	Describe the procedure followed to select HR and HRV indicators (e.g., reviewing previous studies, meta-analyses, etc.). Provide definitions of the indicators and guidelines for their interpretation. Justify the adequacy of the selected indicators to the type of study (e.g., adequate HRV indicators for mental stress).
Ethical issues	Present the procedures to mitigate any potential ethical issue.

Table 7c. Guidelines for reporting HRV research in CTIS: *Data collection*

Topic	Checklist item
HRV software	Mention the software and version used to extract HR and HRV indicators from the data.
Artifact correction	Mention the artifact cleaning methods employed. Report the percentage of corrected beats.
Statistical analysis	Describe the statistical procedures used. Report data transformations. Report and interpret effect sizes in relation to the minimum effect size of interest selected for the sample size determination.

Table 7d. Guidelines for reporting HRV research in CTIS: *Data analysis and cleaning*

5. Concluding remarks

This was our first attempt to collect and use heart rate (HR) and heart rate variability (HRV) as indicators of stress when simultaneously interpreting in different conditions (onsite and remote). The results in this observational, exploratory study showed no difference in stress in our two conditions and did not reach statistical significance. Nevertheless, conducting this study allowed us to refine our methodological design. In spite of the exploratory nature of our study and the drawbacks of a series of methodological limitations and constraints, we believe that this topic merits further investigation. We decided to share what we have learned as a set of methodological recommendations that we hope will be helpful for CTIS scholars wishing to introduce HR and HRV measures in their designs. We have found a way to the interpreter's heart, and now the time has come to explore it.

References

Baghi H. / Khoshsaligheh M. (2019) "Stress in written and sight translation in training setting", *Hikma* 18/2, 237–255.

Bayer-Hohenwarter G. (2009) "Methodological reflections on the experimental design of time-pressure studies", *Across Languages and Cultures* 10/2, 193–206.

Bienertova-Vasku J. / Lenart P. / Scheringer M. (2020) "Eustress and distress: neither good nor bad, but rather the same?", *BioEssays* 42, e1900238.

Billman G. E. (2020) "Homeostasis: the underappreciated and far too often ignored central organizing principle of physiology", *Frontiers of Physiology* 11, 1-12.

Bóna J. / Bakti M. (2020) "The effect of cognitive load on temporal and disfluency patterns of speech. Evidence from consecutive interpreting and sight translation", *Target* 32/3, 482-506.

Can Y. S. / Chalabianloo N. / Ekiz D. / Ersoy C. (2019) "Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study", *Sensors* 19/8, 1849.
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", *Biomedical Signal Processing and Control* 18, 370–377.
Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ, Lawrence Erlbaum.
Cooper C. L. / Davies R. / Tung R. L. (1982) "Interpreting stress: sources of job stress among conference interpreters", *Multilingua* 1/2, 97–107.
Courtney J. / Phelan M. (2019) "Translators' experiences of occupational stress and job satisfaction", *Translation & Interpreting* 11/1, 100–113.
Defranq B. / Plevoets K. (2017) "Over-uh-load, filled pauses in compounds as a signal of cognitive load", in M. Russo / C. Bendazzoli / B. Defrancq (eds) *Making Way in Corpus-based Interpreting Studies*, Cham, Springer, 43-64.
De Rooze B. (2003) *La traducción, contra reloj. Consecuencias de la presión por falta de tiempo en el proceso de traducción*. Unpublished PhD Dissertation, University of Granada.
Ebner K. / Singewald N. (2017) "Individual differences in stress susceptibility and stress inhibitory mechanisms", *Current Opinion in Behavioral Sciences* 14, 54–64.
Faul F. / Erdfelder E. / Buchner A. / Lang A.-G. (2009) "Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses", *Behavior Research Methods* 41, 1149-1160.
Fauquet-Alekhine P. / Rouillac L. / Berton J. / Granry J.-C. (2016) "Heart rate vs stress indicator for short term mental stress", *British Journal of Medicine and Medical Research* 17/7, 1–11.
Feng C. / Wang H. / Lu N. / Chen T. / He H. / Lu Y. / Tu X. M. (2014) "Log-transformation and its implications for data analysis", *Shanghai Archives of Psychiatry* 26/2, 105–109.
Gile D. (2009) *Basic Concepts and Models for Interpreter and Translator Training*. Revised Edition. Amsterdam/Philadelphia, John Benjamins.
Gordon A. M. / Berry W. (2021) "A large-scale study of stress, emotions, and blood pressure in daily life using a digital platform", *PNAS* 118/31, e2105573118.
Gósy M. (2007) "Disfluencies and self-monitoring", *Govor* 24, no. 2: 91–110.
Hedges L. V. (2008) "What are effect sizes and why do we need them?", *Child Development Perspectives* 2/3, 167–71.
Hellhammer D. H. / Stone A. A. / Hellhammer J. / Broderick J. E. (2010) "Measuring stress", in F. Koob / M. Le Moal / R. F. Thompson (eds) *Encyclopedia of Behavioral Neuroscience*, 186–191, New York, Academic Press.
Hjortskov N. / Rissén D. / Blangsted A. K. / Fallentin N. / Lundberg U. / Søgaard K. (2004) "The effect of mental stress on heart rate variability and blood pressure during computer work", *European Journal of Applied Physiology* 92/1–2, 84–89.
Hu D. / Gao L. (2022) "Psychological stress level detection based on heartbeat mode", *Applied Sciences* 12, 1409.

Jakobsen A. L. (2005) "Instances of peak performance in translation", *Lebende Sprachen* 50/3, 111–116.
Jensen A. (1999) "Time pressure in translation", in G. Hansen (ed.) *Probing the Process in Translation. Methods and Results,* Copenhagen, Samfundslitteratur, 103–119.
Kim H.G. / Cheon E.J. / Bai D.S. / Lee Y. H. / Koo B.H. (2018) "Stress and heart rate variability: a meta-analysis and review of the literature", *Psychiatry investigation* 15/3, 235–245.
Klonowicz T. (1994) "Putting one's heart into simultaneous interpretation", in S. Lambert / B. Moser-Mercer (eds), *Bridging the Gap: Empirical Research in Simultaneous Interpretation,* Amsterdam/Philadelphia, John Benjamins, 213–224.
Korpal P. (2016) "Interpreting as a stressful activity: physiological measures of stress in simultaneous interpreting", *Poznan Studies in Contemporary Linguistics* 52/2, 297–316.
Korpal P. (2021) "Stress experienced by Polish sworn translators and interpreters", *Perspectives* 29/4, 554-571.
Kurz I. (2003) "Physiological stress during simultaneous interpreting: a comparison of experts and novices", *The Interpreters' Newsletter* 12, 51–67.
Kurz I. (2002) "Physiological stress responses during media and conference interpreting", in G. Garzone / M. Viezzi (eds) *Interpreting in the 21st Century: Challenges and* Opportunities, Amsterdam/Philadelphia, John Benjamins, 195–202.
Kyriakou K. / Resch B. / Sagl G. / Petutschnig A. / Werner C. / Niederseer D. / Liedlgruber M. / Wilhelm F. H. / Osborne T. / Pykett J. (2019) "Detecting moments of stress from measurements of wearable physiological sensors", *Sensors* 19/17, 1–26.
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", *Frontiers in Psychology* 8.
Lakens D. (2022) "Sample size justification", *Collabra: Psychology* 8/1: 33267.
Lipponen J.A. / Tarvainen M.P. (2019) "A robust algorithm for heart rate variability time series artefact correction using novel beat classification", *Journal of Medical Engineering & Technology*, 43/3, 173-181.
Mackintosh J. (2003) "The AIIC workload study", *FORUM* 1/2, 189–214.
Maier M. / Lakens D. (2021) "Justify your alpha: a primer on two practical approaches." Preprint. *PsyArXiv*, <https://doi.org/10.31234/osf.io/ts4r6>.
Matamala A. / Soler Vilageliu O. / Iturregui Gallardo G. / Jankowska A. / Méndez Ulrich J. L. / Serrano Ratera A. (2020) "Electrodermal activity as a measure of emotions in media accessibility research. Methodological considerations", *JoSTrans* 33, 129-151.
Mellinger, C. D. / Hanson T. H. (2017) *Quantitative Research Methods in Translation and Interpreting Studies*, New York, Routledge.
Moneta G. B. (2018) "Cognitive flow", In J. Vonk / T. Shackelford (eds) *Encyclopedia of Animal Cognition and Behavior*, Cham, Springer.

Moser-Mercer B. / Künzli A. / Korac M. (1998) "Prolonged turns in interpreting: effects on quality, physiological and psychological stress (pilot study)", *Interpreting* 3/1, 47–64.
Moser-Mercer B. (2003) "Remote interpreting: Assessment of human factors and performance parameters", <https://aiic.org/document/516/AIICWebzine_Summer2003_3_MOSER-MERCER_Remote_interpreting_Assessment_of_human_factors_and_performance_parameters_Original.pdf&sa=U&ved=2ahUKEwjkzuOLlqj3AhW8g_0HHZDSAXQQFnoECAMQAQ&usg=AOvVaw0Qfq-K1dE34SVGFyyJ6kre>.
Moser-Mercer B. (2005) "Remote interpreting: the crucial role of presence", *Bulletin suisse de linguistique appliquée* 81, 73–97.
Moses Z. B. / Luecken L. J. / Eason J. C. (2007) "Measuring Task-related Changes in Heart Rate Variability", *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2007*, 644-647.
Plevoets K. / Defranq B. (2016) "The effect of informational load on disfluencies in interpreting. A corpus-based regression analysis", *Translation and Interpreting Studies* 11/2, 202-224.
Pradas Macías M. (2006) "Probing quality criteria in simultaneous interpreting: the role of silent pauses in fluency", *Interpreting* 8/1, 25-43.
Quintana D. S. (2017) "Statistical considerations for reporting and planning heart rate variability case-control studies: reporting heart rate variability studies", *Psychophysiology* 54/3, 344–349.
Quintana D. S. / Alvares G. A. / Heathers J. A. J. (2016) "Guidelines for reporting articles on psychiatry and heart rate variability (GRAPH): recommendations to advance research communication", *Translational Psychiatry* 6/5, e803–e803.
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", *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", *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 19.
Rojo López A. M. / Ramos Caro M. / 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", *PloS One* 16/9, e0257727.
Rosnow R. L. / Rosenthal R. (2009) "Effect sizes: why, when, and how to use them", *Zeitschrift für Psychologie / Journal of Psychology* 217/1, 6–14.
Roziner I. / Shlesinger M. (2010) "Much ado about something remote: stress and performance in remote interpreting", *Interpreting* 12/2, 214–247.
Sapolsky R. M. (2015) "Stress and the brain: individual variability and the inverted-U", *Nature Neuroscience* 18/10, 1344–1346.
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 variabil-

ity", *Frontiers in Psychology* 5, 1040, <https://doi.org/10.3389%2Ffpsyg.2014.01040>.
Shaffer F. / Ginsberg J.P. (2017) "An overview of heart rate variability metrics and norms", *Frontiers in Public Health* 5, 258, < https://doi.org/10.3389%2Ffpubh.2017.00258>.
Sharmin S. / Špakov O. / Räihä K.-J. / Jakobsen A. L. (2008) "Effects of time pressure and text complexity on translators' fixations", *ETRA '08: Proceedings of the 2008 symposium on Eye tracking research & applications,* Savannah, GA, ACM, 123–126.
Slavich G. M. / Shields G. S. (2018) "Assessing lifetime stress exposure using the stress and adversity inventory for adults (Adult STRAIN): an overview and initial validation", *Psychosomatic Medicine* 80/1, 17–27.
Spielberger C. D. / Gorsuch R. / Lushene R. E. / Vagg P. R. / Jacobs G. A. (1983) *Manual for the State-Trait Anxiety Inventory (Form Y1 – Y2)*, Sunnyvale, CA, Consulting Psychologists Press.
Stranks J. (2005) *Stress at Work: Management and Prevention,* Oxford, Elsevier.
Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996) "Heart Rate Variability", *Eur Heart J* 17/28.
Thompson E. R. (2007) "Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS)", *Journal of Cross-Cultural Psychology* 38/2, 227–242.
Tóth A. (2011) "Speech disfluencies in simultaneous interpreting: a mirror on cognitive processes", *SKASE Journal of Translation and Interpretation* 5/2, 23-31.
VandenBos G. R. (ed.) (2015) *APA Dictionary of Psychology*, 2nd ed. Washington DC, American Psychological Association.
Weng Y. / Zheng B. / Dong Y. (2022) "Time pressure in translation: psychological and physiological measures", *Target,* published online, <https://doi.org/10.1075/target.20148.wen>.
Weng Y. / Zheng B. (2020) "A multi-methodological approach to studying time-pressure in written translation: manipulation and measurement", *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19.
Yerkes R. M. / Dodson J. D. (1908) "The relation of strength of stimulus to rapidity of habit-formation", *Journal of Comparative Neurology and Psychology* 18/5, 459–482.

Appendix - Full data

statistic	measurement	session	Mean HR	Mean RR	RMSSD	pNN50	D^2	HF	LF/HF
mean	baseline	1	63.00	959.20	333.86	72.35	1.19	37.87	1.86
		2	73.80	835.40	276.68	56.65	1.43	39.95	1.61
	onsite	1	69.00	888.20	342.04	68.69	0.51	47.78	1.79
		2	68.60	883.40	367.72	72.53	0.74	43.56	1.32
	recovery	1	64.00	944.80	390.02	79.43	0.84	44.98	1.42
		2	68.20	901.80	338.40	64.44	0.45	42.01	1.43
	remote	1	67.40	894.00	335.40	71.90	0.93	49.88	1.17
		2	68.80	879.00	326.54	69.81	1.21	42.83	1.64
median	baseline	1	62.00	973.00	331.30	70.61	0.19	39.20	1.55
		2	67.00	892.00	309.30	60.51	0.79	39.50	1.53
	onsite	1	64.00	932.00	405.70	81.21	0.00	55.67	0.80
		2	69.00	867.00	358.10	73.11	0.01	42.19	1.37
	recovery	1	65.00	917.00	427.40	86.81	0.00	52.12	0.92
		2	65.00	923.00	370.90	74.90	0.01	39.51	1.52
	remote	1	66.00	913.00	377.50	82.83	0.00	48.05	1.08
		2	68.00	886.00	380.40	75.31	0.27	41.33	1.42
SD	baseline	1	5.43	79.27	92.29	17.18	1.55	10.45	1.05
		2	14.75	133.32	145.63	26.04	1.30	8.77	0.63
	onsite	1	11.79	132.15	176.96	34.20	1.15	19.26	2.21
		2	8.44	106.31	98.71	17.44	1.60	5.65	0.30
	recovery	1	7.07	103.72	110.82	18.37	1.88	13.34	0.83
		2	12.19	139.19	125.14	27.22	0.77	7.15	0.40
	remote	1	5.32	70.90	161.25	24.46	2.08	14.32	0.75
		2	6.61	78.39	121.47	18.41	1.66	15.85	1.07