Testing the impact of remote interpreting settings on interpreter experience and performance

Published Version:
Chmiel, A., Spinolo, N. (2022). Testing the impact of remote interpreting settings on interpreter experience and performance. TRANSLATION, COGNITION & BEHAVIOR, 5(2), 250-274 [10.1075/tcb.00068.chm].

Availability:
This version is available at: https://hdl.handle.net/11585/902724 since: 2026-01-29

Published:
DOI: http://doi.org/10.1075/tcb.00068.chm

(Article begins on next page)

TESTING THE IMPACT OF REMOTE INTERPRETING SETTINGS ON INTERPRETER EXPERIENCE AND PERFORMANCE

Methodological challenges inside the virtual booth

Agnieszka Chmiel & Nicoletta Spinolo

Adam Mickiewicz University in Poznań | Università di Bologna

This paper presents methodological challenges in a study focusing on the impact of remote interpreting settings on interpreter experience and performance. In recent years, the practice of simultaneous interpreting has undergone a robust development with the quick uptake of remote interpreting technologies due to the global pandemic. In order to investigate remote interpreting, we created the Inside the Virtual Booth project encompassing a survey and an experimental study. We report on selected results of the survey that directly inform the experimental study design. We focus on challenges related to the compromise between experimental control and ecological validity, creation of materials and selection of dependent variables, including eyetracking measures that cannot be directly applied from reading studies to a study involving multimodal content typical for remote interpreting assignments. The paper may serve as a source of methodological guidance to scholars entering the field of experimental translation and interpreting studies.

Keywords: remote simultaneous interpreting, methodology, eyetracking, interpreting studies, virtual booth

1. INTRODUCTION

With the strong development of Cognitive Translation and Interpreting Studies (Halverson & Muñoz Martín 2021) and Translation Process Research – TPR (Jakobsen 2017), research methodology has attracted more attention from translation and interpreting scholars within recent years. This is also evidenced by a growing number of publications offering assistance to researchers in this area (Saldanha & O'Brien 2014; Angelelli & Baer 2016; Mellinger & Hanson 2017). On the other hand, recent developments in interpreting practice have created new methodological challenges. The introduction of technology in the conference interpreting workflow is changing the professional landscape at an unprecedented speed (Fantinuoli 2019), with an even greater impact of the global pandemic on the fast uptake of Remote Simultaneous Interpreting (RSI) platforms and tools.

This paper focuses on the methodology of a research project entitled Inside the virtual booth: the impact of remote interpreting settings on interpreter experience and performance. The project includes two stages, a survey and an experimental study, and aims to benefit from a process-oriented approach to the investigation of cognitive aspects of RSI. It entails a multi-method, multi-source data collection process that inevitably involves methodological issues that needed to be addressed in terms of experiment and material design, preparation and tools. In this paper, we start by offering selected survey results and show how they inform decisions in experimental design. Furthermore, methodological considerations are made in terms of challenges related to study design and compromise between experimental control and ecological validity, proper control of experimental materials and selection of appropriate dependent variables. These include eyetracking measures that cannot be directly applied from reading studies to a study involving multimodal content typical of remote interpreting assignments. Thus, the paper may also serve as a source of methodological guidance to scholars entering the field of experimental translation and interpreting studies.

2. THE 'INSIDE THE VIRTUAL BOOTH' PROJECT

The project was funded by AIIC, the International Association of Conference Interpreters, as the first AIIC Research Grant, set up by the association in 2020 to foster its connection with the interpreting studies community and to fund studies that could be of interest for AIIC members. The project aims at studying the impact of RSI on both interpreters' experience (understood here as user experience, i.e. the interpreter's preferences and views when performing RSI, rather than as interpreting expertise) and performance.

Despite starting to develop well before the COVID-19 pandemic (Moser-Mercer 2005; Mouzourakis 2006; Roziner & Shlesinger 2010), RSI has gained incredible momentum during it, so much so that it seems to have become a standard for conference interpreters, with significant impact on the practice of simultaneous interpreting and potentially game-changing cognitive implications in terms of human-computer interaction (using a soft console instead of the traditional hard one) and also human-human interaction (with boothmate, other colleagues and, in general, participants in the event) (Mellinger 2019). However, this rapid expansion has made it hard for research to keep up. As a matter of fact, at the time of writing, various research endeavours are ongoing and aim at studying different aspects of RSI, but very few results have been published yet. Seminal research carried out on remote interpreting through hard consoles and traditional booths found that remote interpreting can be subjectively perceived as more stressful and tiring by interpreters, although objective measurements did not seem to point in the same direction (Moser-Mercer 2005; Roziner & Shlesinger 2010). The studies also showed that the feeling of 'lack of control' (Moser-Mercer 2005, 94) and presence can be perceived as a consequence of not being physically present in the communicative event (see also Seeber et al 2019). As far as RSI with soft consoles is concerned, research has been carried out on interpreter preferences regarding platform interface design (Saeed et al 2022) and on the integration of automatic speech recognition in RSI platforms (Defrancq & Fantinuoli 2021; Fantinuoli et al 2022). In order to contribute to the yet small body of research on RSI, this project aims at gaining insight into the impact of different RSI configurations (regarding both boothmate location and technical setup) by answering the following research questions:

1. How do remote interpreting settings influence user experience, interpreter's stress levels, mental workload and performance?

2. Is there a difference in user experience and interpreting performance between interpreters working in hubs with their boothmate on-site and remotely in virtual booths?

3. To what extent is the interpreters' performance and experience affected by different types of multimodal input?

The project has a two-stage design. Stage one envisaged a survey aimed at obtaining detailed feedback from professional interpreters on their use of RSI platforms and on their user experience, with the goal of informing experimental design for stage two of the project. Stage two envisages an experimental study of different configurations of RSI and boothmate interaction examined through eyetracking data, performance data and questionnaires on user experience, anxiety and subjective mental workload. This paper briefly reports on stage one and then focuses on methodological choices made in experimental design.

3. Stage one: Survey

The survey was developed using LimeSurvey (2003) and disseminated through AIIC channels (newsletter, social networks) and through the researchers' contacts and social networks. In the next sections, we discuss its design and piloting and present selected results that justified our choices in the experimental design.

3.1 Design and piloting

The survey included 39 questions divided into 6 main sections: *Demographics, Experience with RSI, User experience, Setting, Equipment, Best practices*. The Demographics section collected data on the population surveyed, such as age, gender, working language(s), country of residence, years of expertise, professional status (in-house/freelance), association with AIIC (because of the expertise it ensures), conference days worked in the previous six months. *The Experience with RSI section* asked respondents if they had ever worked in RSI, if they did before the pandemic and how often they had worked in RSI in the previous six months. It also asked them about RSI training and how frequently they used different RSI platforms. The User experience section collected more detailed information about the respondents' experience with the two platforms they used most often, focusing on their pragmatic qualities.

The *Setting* section investigated the frequency (on a 5-point Likert scale) with which respondents worked from home or a personal office or from a hub and with different boothmate locations: co-located, not co-located, no boothmate. The Equipment section aimed at gaining an insight into respondents' setup when working with RSI in terms of additional devices, screens and connection in the two different settings (home and hub). Finally, the Best Practices section aimed at further understanding the interaction with event stakeholders: technical assistance, boothmate interaction and interaction with event participants (event chat, Q&A formats).

The questionnaire was piloted with a group of eight interpreters who filled in the survey once and then again after two weeks to check for test-retest reliability. A test-retest reliability analysis was performed (following Mellinger & Hanson 2017) on all questions including scales (frequency and semantic differentials) with a resulting Spearman's correlation coefficient of 0.806. The interpreters participating in the pilot also provided useful feedback; for example, one of them suggested to ask the average number of days worked over the previous 6-months instead of 12, because the first part of 2020 had been an anomalous one for everyone in the market, while another tester suggested not to refer to Zoom as an RSI platform, as it is not specifically designed for RSI.

3.2 Demographics

The survey had a total of 385 responses, out of which 295 were complete and 90 were partial; 222 came from AIIC members. On its website AIIC reports the total number of members to be 3049; therefore, the required sample size for statistical representativeness would be 342 (95% confidence level and 5% error); however, as explained in the introductory section to the survey, it was directed only to interpreters who had already had experience in RSI: it was thus impossible to know the exact size of the population we were investigating and the required sample size.

In terms of age, 6.51% respondents were younger than 30, 21.09% were between 31 and 40, 28.65% between 41 and 50, 21.09% between 51 and 60 and 20.05% older than 60; 2.6% decided not to answer the question on age (N=384). Regarding gender, 67.19% were female, 27.6% male, 0.78% selected 'other' and 4.42% did not answer or preferred not to say. The respondents' geographical distribution is reported based on AIIC's regions: most of them (N=371) were based in Europe (63.07%), followed by North America (12.4%), Central and South America (11.86%), Africa and Arab Countries (7.01%), and Asia-Pacific (5.66%).

3.3 Expertise

The sample was quite varied as regards professional expertise measured in years: 26.01% of respondents had fewer than 10 years of working experience; 26.81% had 11 to 20 years; 27.08% had 21 to 30; 12.06% had 31 to 40 and 8.04% had more than 40 (N=373). 59.95% (N=372) respondents were members of AIIC and, of these, 91.93% were active members, which means they have at least 150 days of working experience. The 40.05% who had no association with AIIC were also experienced professionals as they had worked an average of 46 days in the previous six months (SD 36.36).

Interestingly, most respondents were fairly new to RSI, as 73.5% had not worked in RSI before the pandemic, while 26.5% had. They had, however, quickly become frequent users of RSI platforms, with more than 44% of them using these more than once a week and 11.2% every day (Figure 1).

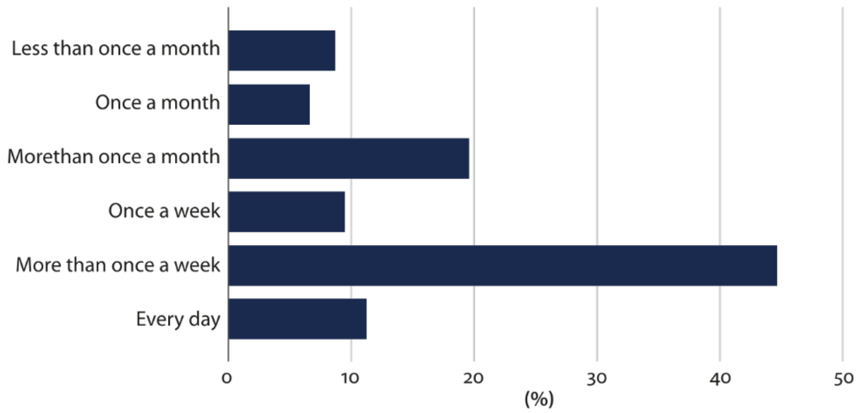

Figure 1. Frequency of work in RSI mode in the previous 6 months (N=347)

3.4 Selected results: Insights for the experimental stage

In this section, we present selected results from the survey that have been particularly useful to inform the experimental design proposed in Section 3.

In the *Experience with RSI* section of the survey, and in order to select the platforms for the experiment, participants were asked to indicate their first and second most frequently used platforms; results showed that Zoom was the most frequently used one, with 78.5% of 311 respondents indicating it as their first choice, while the second most frequently used one by most respondents (29.6%; N=203) was KUDO. Respondents were asked to rate how often they used different platforms (the list proposed was based on Collard & Buján 2021) on a 5-point Likert scale and results, in line with Collard and Buján (2021), showed that Zoom with interpretation feature was overwhelmingly the most used (32.95% use it always and 44.8% use it often; N=346). This led us to select this platform for our experiment.

As per our data from the *Setting* section of the survey, most interpreters work in RSI from home (33.7% always and 40.9% often), although a fair share also work from hubs (3.3% always, 13.5% often, 23.8% sometimes). Participants were also asked to indicate how often they worked with no boothmate, a co-located one (i.e. one with whom they shared physical space) or a not co-located (i.e. remotely located) one, with the latter condition being the most frequent (Figure 2).

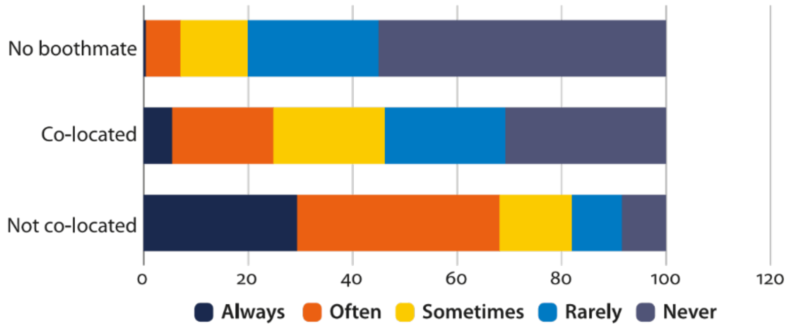

Figure 2. Boothmate location for survey respondents, % (N=304)

Since the most frequently identified setting was *working from home with a not co-located boothmate*, it becomes very important to look at how participants usually communicate with their partner. Figure 3 (Best Practices) illustrates the responses, showing the external chat on a separate device as the most frequent option.

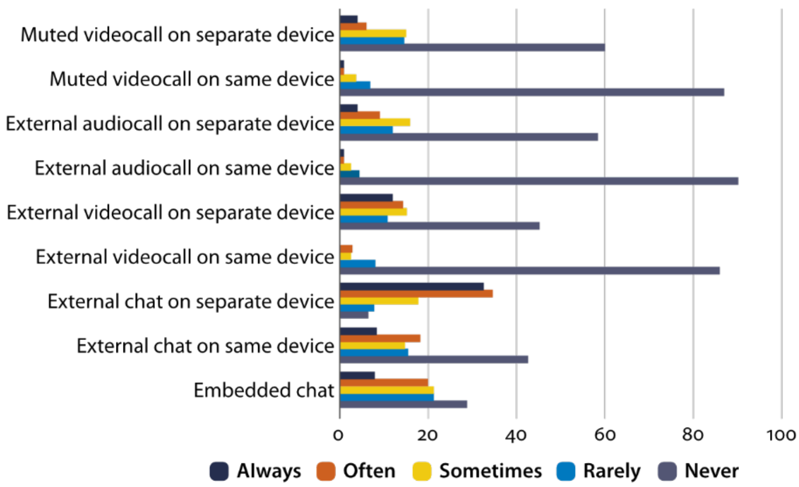

Figure 3. Interaction with boothmate (N=300)

In terms of visuals during presentations, as Figure 4 shows, most interpreters are used to seeing both the speaker and the slides.

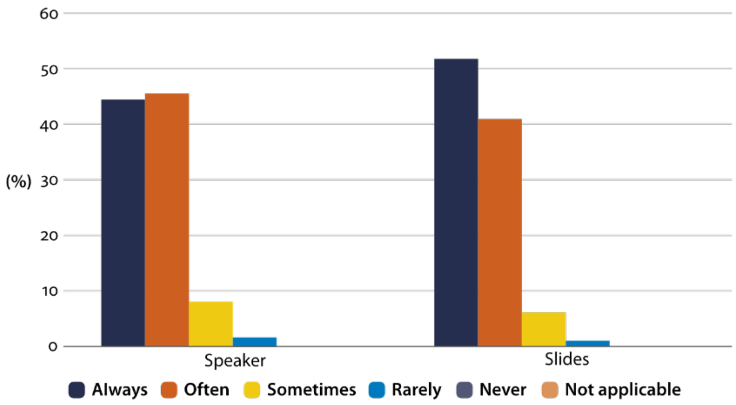

Figure 4. Visuals during presentation (N=298)

In RSI, question and answer (Q&A) sessions can be particularly complex, as they might be asked through the event's chat or aloud, while questions may or may not be read aloud by a speaker or moderator. As Figure 5 illustrates, interpreters can in most cases see and hear delegates asking questions, while it often happens that the questions are written and read out. In some cases, however, interpreters can hear but not see delegates asking questions, or questions are asked only in writing and not read out.

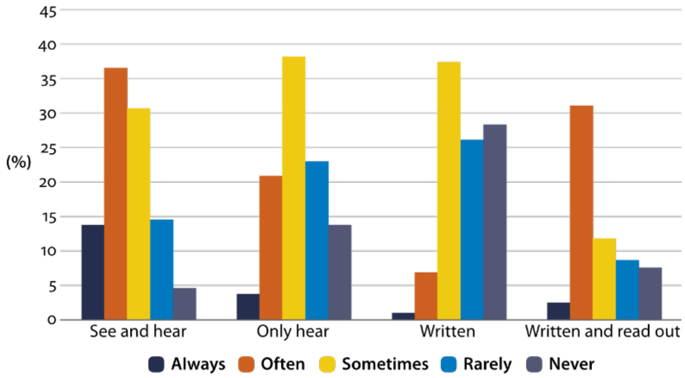

Figure 5. How is Q&A usually organized? (N=290)

The survey has helped identify certain characteristics of RSI settings that are worth investigating in the experimental study: these include working individually with a boothmate that is not co-located with the active interpreter. This means that the communication with the boothmate is impaired and frequently achieved through a chat only. The interpreter thus receives various multimodal inputs, such as the speaker's video, slides, the delegates' video and/or audio during Q&A, the boothmate's communication via chat and the delegates' questions via chat. We thus focused on these aspects when designing the experiment described below.

4. Stage two: Experimental study

Insights gained through the survey informed the design of the experiment aiming to investigate the impact of various remote interpreting settings on the interpreter's experience and performance. Care was taken to achieve the ecological validity of the experiment, understood in line with Mellinger & Hanson (2022) as pertaining to the setting, the stimuli and the task.

4.1 Study design

The experiment is planned as a staged experiment with a confederate, i.e. a research actor to play the part of the boothmate. Such design is rarely found in Interpreting Studies: some studies on dialogue interpreting include actors who act out a dialogue to be interpreted in an experimental task by an interpreter participating in the experiment (Hale et al 2022; Tiselius & Sneed 2020). This is to obtain ecological validity since dialogue interpreting involves a lot of interaction and would be difficult to investigate on the basis of recordings. In such cases, actors follow a pre-prepared script and are asked to use specific problem triggers as instructed. However, for the interpreting situation to be viable, they have to follow the interpreter's output if the interpretation is not correct and deviates from the script. They are then asked to return to the script whenever possible. This is a required compromise – the input text differs slightly for participants but the setting and the dialogue are realistic. To the best of our knowledge, only a single study to date has used a staged setting for the investigation of simultaneous interpreting (Seubert 2019). Since simultaneous interpreting is more monologic and involves less interaction, deviations from the script are not required in response to the interpreter's production. In our case, the parts of the speaker, the delegates asking questions and the boothmate will be played by research actors to increase ecological validity. The boothmate will be instructed to offer only the assistance planned in the script even though this might not align with previous agreements with the investigated interpreter – this is to ensure comparability of data across participants. Obviously, if the boothmate's assistance does not meet the interpreter's expectations, this might affect the interpreter's perceived stress. Unfortunately, this is a confounding variable that cannot be controlled for the sake of ecological validity.

Since the most frequently used platform identified in the survey stage was Zoom, we decided to use it, even though it is not a specialised RSI platform. It will be used with different boothmate presence settings leading to three experimental conditions: a co-located boothmate, a not co-located boothmate that would communicate with the active interpreter through a chat channel, and a not co-located boothmate in a virtual booth. A virtual booth is understood here as a backchannel set up on the same device for communication between boothmates who can see and hear each other, thus communicate in an oral or written form and easily achieve a successful handover. The virtual booth will be set up through GT Booth (2022), a solution that is currently offered for free to interpreters using it as a backchannel. To offset fatigue and task order effects, the sequence of the conditions will be counterbalanced across participants using a full Latin square design (Bradley 1958). The conditions selected for the experiment are based on the survey results: we selected the most frequently used platform and the most frequently used mode of communication with the not co-located boothmate. The third condition including the virtual booth was selected in order to test a solution that tries to bridge the communication gap between not co-located remote simultaneous interpreters. Thus, the first independent variable in the study (boothmate presence) has three levels: a co-located boothmate, a boothmate connected via chat and a boothmate in a virtual booth.

The second independent variable is the type of multimodal input. Each text will include four different types of input: audio and video of the speaker (AV condition); audio and video of the speaker and slides (AV-slides condition); text input in the chat section (chat condition); audio input from the floor (audio condition). These conditions are not fully counterbalanced in terms of content or sequence due to ecological validity issues. It is typical for a conference to first include presentations that are later followed by questions and comments from the floor – either spoken or written in the chat or Q&A box. Also, the content of the presentation cannot be plausibly exchanged with the content of the Q&A session. The AV and AV-slides conditions will always precede the chat and audio conditions for ecological validity.

We will test two groups of participants with two different language combinations. Interpreting will always be performed from B (or C) language to A. The two combinations are English-Polish and Spanish-Italian, which means that all the materials have to be prepared for two language versions since we will be testing participants with two different language combinations. This has both advantages and disadvantages. The upside is that we will be able to obtain results that are language-pair independent. Many studies that apply only one combination run the risk of not being able to generalise across other language pairs. However, such a design poses an additional challenge of preparing comparable materials, an issue we discuss in the materials section below.

4.1.1 Participants

The participants include 32 professional interpreters working in two language combinations (16 with Polish A and English as B or C and 16 with Italian A and Spanish B or C). An attempt will be made to invite as many AIIC members as possible as this ensures a high level of expertise.

The participants will self-report proficiency in four language skills (reading, writing, speaking, listening) for the languages in their combinations. If not AIIC active members, they will also self-report their professional experience measured as years of experience, total estimated number of conference days, average number of conference days within three years pre-pandemic (2017–2019) and average number of conference days during the pandemic. This differentiation is introduced because the pandemic has changed the way many interpreters work and the pandemic period might not be truly representative of their exposure to professional assignments.

All professionals will answer a short demographic questionnaire and a series of questions about their experience with RSI (taken from Stage 1 survey). Specific questions will be asked about the participants' familiarity with and previous exposure to the technological solutions used in the experiment (Zoom and GT Booth) to control for the confounding familiarity effect. Additionally, the participants will complete the Technology Adoption Propensity Index (Ratchford & Barnhart 2012) as adapted for interpreters by Mellinger & Hanson (2018) to further gauge the participants' position to technology.

The participants will be recruited through personal contacts and social media. They will be remunerated for their participation in the experimental session.

4.1.2 Materials: Challenges when matching experimental stimuli

Materials for each language pair include three 15-minute speeches. In order to investigate the actual influence of the manipulated variables on the interpreters' behaviour, the confounding variable of text difficulty has to be controlled. Thus, we need to use three texts matched for difficulty on a variety of metrics. The most obvious ones are length (number of words) and source text delivery rate (words per minute and syllables per minute as the latter measure is more comparable across languages). High delivery rates have been shown to negatively influence interpreting performance (Korpal & Stachowiak-Szymczak 2019; Plevoets & Defrancq 2016). We thus aimed at 120–130 words per minute as a moderately easy delivery rate (Setton & Dawrant 2016). Although professional interpreters frequently handle faster input (Seeber 2017), we favour a moderate speed value in order not to create ceiling effects (i.e. the task being too easy and the resulting performance measures are close to the maximum score) (Chmiel 2021a) or floor effects (where the task is too difficult and performance measures are close to the minimum).

Text difficulty can be matched using other both subjective and objective criteria. Seeber et al (2020) used naturalistic materials and controlled the critical phrases for mean duration, mean number of words, mean number of letters per word and probability of fixation. The last two measures were taken into account as this was an eyetracking study focusing on reading the text. Other scholars using authentic materials asked professional interpreters or interpreting trainers to subjectively judge text difficulty (Díaz-Galaz et al 2015; Shao & Chai 2020). However, to better control for difficulty, many scholars manipulate authentic materials to make them more comparable. For instance, Gieshoff (2018) omitted potential problem triggers such as numbers and proper names and reformulated long sentences to include only sentences with no more than one subordinate clause. Su & Li (2019) matched their texts on the number of words, lexical density and the number of complex syntactic constructions (such as embedded clauses). In order to obtain a holistic objective measure of text difficulty, a readability index may be used (Chmiel & Mazur 2013; Lin et al 2018; Liu et al 2004). Readability indices usually measure text difficulty on the basis of word and sentence length.

Our experimental texts will be matched for length (approx. 1800 words) and readability score measured with the Gunning Fog index (Gunning 1969) while the actual recordings will also be matched for speed. They will also include an even number of evenly distributed problem triggers, including numbers, proper names and low-frequency nouns. These have been found to increase cognitive load (Gile 2009), are frequently subject to boothmate assistance (Chmiel 2008) and will be used to measure interpreting accuracy. Since complexity of numbers influences their difficulty (Defrancq & Fantinuoli 2021; Jones 2002), numbers will be matched for difficulty based on the classification used by Kajzer-Wietrzny et al (2021) and extended by Rozkrut (2022). Proper names will be matched by type (an even number of surnames, placenames, etc.). Low-frequency nouns will be matched for frequency based on SUBTLEX-UK (Van Heuven et al 2014) and SUBTLEX-ESP (Cuetos et al 2011).

The materials will first be developed in English for the English-Polish group of participants. They will then be translated into Spanish for the Spanish-Italian group of participants. Translation will be done by a professional English-Spanish translator. A similar procedure was adopted in another study to produce various language versions of the experimental stimuli (Hale et al 2022). If need arises, the translated versions will be adjusted so that they meet the matching criteria (length and readability score). They will be checked for naturalness by a native speaker of Spanish. Low-frequency nouns will be translated using translation equivalents with matching frequency in Spanish.

The texts will be recorded with the speed matching that for the English texts. Some studies report differences in average speaking rates among native speakers of Spanish and English: Spanish speakers tend to speak faster than English ones (see Seeber 2017), which might mean that if the English and Spanish source texts have a similar speed, the Spanish ones might appear easier to the SpanishItalian interpreters than the English ones to the English-Polish interpreters since Spanish-Italian interpreters might be more accustomed to higher speaking rates. However, we prefer having the texts with matched absolute speaking rate values at the expense of leaving the subjective perceptions uncontrolled for.

The choice of topic is also an important issue to be considered when designing experimental stimuli. The topic should not be too narrow for a study that does not involve prior preparation as preparation has been shown to affect performance (Díaz-Galaz et al 2015). On the other hand, it should not be too general because our stimuli should include low-frequency nouns. Some study participants may be familiar with a given topic so it is important to inquire about this after the experiment to control for the confounding familiarly effect. When selecting a topic for creation of two or more matched speeches, it is useful to search for suitable source materials to base the speeches on. In our case, we selected country reports for the Environmental Implementation Review prepared by the European Commission (Environmental Implementation Review 2022). These are similarly structured documents that present information on the implementation of the EU environmental policies. The documents include some terminology and many statistics. We used the data from the reports to build conference speeches around them.

4.2 Dependent variables

The project envisages collection of various types of data to examine processing, performance and experience including preferences, stress and workload levels. We will employ both objective and subjective measures. The objective ones include eyetracking and accuracy measures while the subjective ones will be based on questionnaires.

4.2.1 Eyetracking measures as an index of processing in RSI

Eyetracking studies of interpreting, which have become more numerous in the last decade, have so far largely drawn from reading research and used predominantly reading measures (Chmiel 2021b). The method is frequently used in studies involving sight translation and simultaneous interpreting with text. They either compare sight translation to reading (Dragsted & Hansen 2009; Huang 2011), focus on directionality (Su & Li 2019), or investigate various problem triggers, such as syntax (Chmiel & Lijewska 2019; Shreve et al 2010) or numbers (Chmiel et al 2020; Korpal & Stachowiak-Szymczak 2018; Seeber 2012; Stachowiak 2016). So far, no studies using eyetracking in remote interpreting have been published although we are aware of several projects that are underway. Since our stimuli will be multimodal, we have to rely on other than just text reading measures, especially as verbal input in the form of written text will be visible only on slides and in the chatbox in our design.

There are few eyetracking studies involving multimodal input in interpreting. Seeber and Kerzel (2011) and Gieshoff (2018) used pupil dilation as an index of cognitive load in a simultaneous interpreting task. Stachowiak (2016) presented her participants with congruent and incongruent graphic visuals during a simultaneous interpreting task – she used fixation durations and counts as her measures of cognitive effort. Prandi (2022) used time to first fixation, average fixation duration and total fixation time in her study of term search behaviour when using various tools (traditional glossaries, computer-assisted interpreting tools and automatically generated translations of terms). An eyetracking study that has used the largest variety of stimuli so far is by Seubert (2019), who eyetracked interpreters working at a simulated conference with various events, such as slides with varied visuals, using a laser pointer by the speaker, people entering the conference room and walking past the booth, the speaker approaching the listeners, etc. She analysed her eyetracking data by creating various areas of interest (for instance including the slide, the speaker, the public) and calculated fixation count and duration. For different slide types she analysed the interpreters' division of visual attention among the speaker and the slides via total viewing time. For some slides, a delay between the slide presentation and the first fixation on the slide was calculated. She also presented gaze plots and gaze point clusters for more dynamic events (such as the speaker approaching the audience). Gaze plots visually depict the location, the sequence and the length of fixations. These are series of connected circles reflecting fixations with numbers depicting consecutive fixations and circle size growing with fixation length. Gaze point clusters visually present areas in which several fixations are located within a given period (Seubert 2019). However, these were only analysed qualitatively.

Taken together, these studies have used either traditional measures (such as pupil dilation, fixation count and duration) or measures better suited for complex eyetracking of dynamic environments (gaze plots and gaze point clusters) – but qualitatively only. We thus had to look for additional measures in other areas of research.

The first obvious choice was to look at neighbourhood areas within Translation Studies, i.e. studies on written and audiovisual translation. In eyetracking studies investigating written translation, the usual setup includes two separate windows for the source text and the target text. Apart from usual measures (such as fixation duration and count), shifts of visual attention between source text and target text areas are also measured (Jakobsen & Jensen 2008; Hoberg 2022). Studies on the integration of subtitle reading and film watching have used a parallel measure called a revisit, defined as a number of glances made back to the area of interest (a subtitle in this case) after visiting it for the first time (Doherty & Kruger 2018; Szarkowska & Gerber-Morón 2019). It is used as an indicator of processing difficulty for that area. We believe that attention shifts between various interest areas in our study (such as the speaker, the slides, the chatbox) and, potentially, revisits to these areas seem to be promising eyetracking measures for investigating processing during RSI.

We then turned to other disciplines and studies on visual processing, cognitive load and fatigue and found promising measures that could be applied to our study: Index of Pupillary Activity (IPA) and mental fatigue and concentration measures based on fixations and visual attention entropy (i.e. distribution and randomness of fixations).

IPA is a pupillometric measure that indexes cognitive load on the basis of pupil diameter oscillations (Duchowski et al 2018). Usually, measuring the pupil is problematic due to its sensitivity to light. IPA is based on computations that focus on relative oscillations and thus offset the confounding effect of illumination (Duchowski et al 2018). Thus, IPA seems a promising measure to be analysed in our study with varied luminosity of our dynamically changing experimental stimuli.

A study of fatigue of airplane pilots (Naeeri et al 2019) shows that fixation count, fixation duration and the visual attention entropy measure can be used as indicators of fatigue and concentration. With increasing fatigue and decreasing concentration, mean fixation count decreases while mean fixation duration entropy increase. High entropy value reflects random and non-focused distribution of visual attention and could be potentially used in our study to measure the distribution of visual attention when interpreting remotely in various conditions.

We will collect eyetracking data with an EyeLink table-mounted eyetracker (SR Research, Ontario – Canada) used in the remote mode, i.e. without asking the participant to keep the head on the forehead rest to minimize head movement. Its recording frequency is 500 Hz, which means 500 measurements of eye locations per second. Since our fixation-based measures will be calculated for areas of interest larger than a single word (such as a chat window, a slide view in Zoom), other eyetrackers with lower frequency values could be potentially usable in this case, as well. However, data collection and analysis should always be piloted when using lower frequencies for eye movement recording to see if data quality is sufficient for the analysis planned.

4.2.2 Accuracy as an index of performance in RSI

Assessing the quality of interpreting is a daunting task since various assessment methods, be it propositional accuracy score (Bartłomiejczyk 2006) or grading scales (Tiselius 2009) have their limitations, including subjectivity and graders' workload leading to fatigue. We thus decided to use translation accuracy of the problem triggers included in experimental materials as an index of performance.

The disadvantage is that only selected parts of the interpreters' performance will be assessed, but the advantage is that the assessment may be much more precise and objective. A similar approach is frequently adopted by other scholars (Liu et al 2004; Defrancq & Fantinuoli 2021). If accuracy assessment criteria are sufficiently specific and objective, single assessment of each item should be reliable. Fortunately, our problem triggers lend themselves to setting such objective criteria much easier than longer textual units, sentences or passages.

The accuracy of interpreting proper names will be assessed by comparing the source text version to the target text version. The interpretation will be assessed as correct (if the proper name is fully repeated), as an approximation (if it is repeated in a slightly changed form, e.g. Kokkoma instead of Kokkola or SustainBank instead of SustainBaltic), inaccurate or omitted. Low-frequency nouns will be assessed as interpreted correctly, if the appropriate equivalent is provided (e.g. rosomak as a Polish equivalent of wolverine), interpreted incorrectly or omitted.

Numbers might be more problematic in accuracy judgements because infelicities might be related to the specific digits while maintaining the correct order of magnitude or vice versa. Numbers may also be approximated and these approximations might also vary depending on the number complexity. We are planning to adopt the accuracy grading categories based on Kajzer-Wietrzny et al. (2021) and adapt it if needed to include more fine-grained categories. The categories are: completely accurate, good approximation (e.g. 1995 is rendered as the nineties), omission, completely inaccurate (the rendition includes a specific value that differs from the source text one).

Care will be taken to include problem triggers in non-redundant contexts so that omissions of these problem triggers cannot be justified. Such specific delineation of accuracy grading criteria should increase the objectivity of accuracy judgements and reliability of assessments.

4.2.3 Questionnaire data as indices of subjective experience in RSI

Participants' subjective experience with RSI will be assessed through different self-reported measures. Their user experience will be assessed with a reduced version of the UEQ (User Experience Questionnaire; Laugwitz et al 2008), a validated user experience questionnaire containing six scales and a total of 26 items. As the questionnaire's manual explains (Schrepp 2019) it is possible to select some scales out of the questionnaire, depending on the researcher's needs and objectives. We will only use the four scales that assess pragmatic qualities (Attractiveness, Efficiency, Perspicuity, Dependability). More specifically, Attractiveness measures the overall impression users have of the product; Efficiency means to what extent users can solve problems without unnecessary effort; Perspicuity focuses on how easy it is to get familiar with the software, while Dependability reports to what extent the user feels in control of the interaction.

The UEQ has already been used successfully in translation and interpreting studies by Braun et al (2020) to assess two different kinds of Virtual Learning Environments for interpreter training. Participants will be asked to fill in the UEQ after each text, so as to complete it with a recent use of the technological solutions in mind (Schrepp 2019). Participants should be able to complete the questionnaire in 3–4 minutes, and results will reflect the participants' perceptions when using the technological solutions applied in the study. Participants' previous familiarity with Zoom and GT Booth (i.e., their being accustomed to using them, self-reported before the study) will be taken into consideration as a potential confounding variable and thus will be controlled for in the subsequent statistical analysis when extracting results.

The interpreters' perceived level of anxiety while performing the different tasks will be measured through the STAI (State-Trait Anxiety Inventory; Spielberger et al 1983). The STAI aims at measuring state anxiety, determined by a specific situation the subject is experiencing (form Y-1), and trait anxiety, which is defined as a characteristic of a subject's personality (form Y-2). The STAI has already been successfully used in translation and interpreting studies; for example, Moser-Mercer (2005) used it together with other physiological and selfreported measurements to compare feelings of anxiety generated by interpreting in the on-site and remote conditions; she administered the Y-2 form weeks before the experiment, to measure her subjects' trait anxiety, and then the Y-1 form after each session in the study and found no statistically significant difference with the remote condition. Korpal (2017) used it, in combination with other linguistic and psychological indicators, to study stress in simultaneous interpreting, and measured anxiety before and after simultaneous interpreting tasks. Rojo et al (2021a), on the other hand, studied the influence of time-pressure on translation students and used the STAI, again coupled with other physiological and self-reported measures, to study stress and anxiety. Rojo et al (2021a) asked their participants to fill in the STAI (forms Y-2 and Y-1) after the first task and then to again fill in the STAI Y-1 after the second task. Rojo et al (2021b) also used the STAI, together with other parameters, to measure stress and anxiety in interpreting students during a dialogue interpreting session; in this case, they asked subjects to complete the STAI (both forms) before the task and then the STAI Y-1 after the task. We will have subjects complete the STAI Y-2 and STAI Y-1 (as a baseline state measurement) at the beginning of the experiment, before any task, and then the STAI Y-1 after interpreting each of the three experimental texts.

Participants' subjective mental load will be measured through another validated tool, the NASA-TLX (Hart & Staveland 1988). As the NASA-TLX manual states, the tool 'provides an overall workload score based on a weighted average of ratings of six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort and Frustration' (NASA-TLX 1986:1). The NASA-TLX has been used in Translation and Interpreting Studies; for example, Sun & Shreve (2014) used it to assess translation students' perceived level of difficulty of translation texts while Liu et al (2019) and Yang et al (2022) used it in combination with eyetracking measurements and retrospection to assess translation difficulty. Our participants will be asked to complete the NASA-TLX three times, after each experimental condition, so as to compare subjective mental workload when working with a co-located boothmate, a not co-located boothmate and a boothmate located in a virtual booth.

5. CONCLUSIONS

We reported on methodological challenges involved in a project that uses both survey and an experiment as research methods. We believe that with the increasing awareness of the translation and interpreting studies community as to the importance of methodological rigour, this paper can offer useful insights and guidance.

The project made it possible to create a synergy effect between the survey stage and the experimental study because the survey results directly informed our choices regarding the manipulation of independent variables in the experiment (the platform used, the settings for cooperation with a boothmate to be used in the experiment, etc.). Most of our predictions were confirmed but it is much safer to base a study on hard data rather than personal conjectures. This is especially true in this case since the topic of RSI is still new and under-researched and it is thus difficult to find justification for experimental manipulations in existing literature.

As in any experimental study, choices have to be made to reduce the complexity of the investigated task in order to control for selected variables. In our study of remote interpreting, we decided to focus on differences in boothmate presence and various multimodal input channels. Boothmate presence was operationalised through different settings in which a boothmate can offer assistance (a traditional co-located boothmate, a not co-located boothmate using only chat as a communication channel and a not co-located boothmate using a virtual booth). We included four input channels most frequently occurring in an RSI setting (audio and video of the speaker, audio and video of the speaker and slides, text input in the chat section and audio input from the floor). We did not include a manipulation of another important variable in RSI, the sound quality. It was a compromise to be made – too many variables would make the design too complex and the findings too noisy. Additionally, limitations not predictable at the design stage might emerge during data collection and will have to be considered when interpreting the results.

Our dependent variables also had to be selected carefully. We believe that using novel eyetracking measures (such as the Index of Pupillary Activity and the visual attention entropy), alongside more traditional reading-based ones, will offer interesting insights for RSI settings we are investigating. Together with interpreting accuracy as an objective measure of performance and with subjective measures of user experience, anxiety and mental workload, these measures will provide interesting and multifaceted findings on interpreters' RSI experience and performance.

We hope that detailed descriptions of all the considerations and decisions made when selecting variables and matching experimental stimuli, as well as the synergy between a survey and an experiment we have benefited from will serve as a source of guidance to scholars planning their first experiments in the field of cognitive translation and interpreting studies.

FUNDING

This research has been funded by the AIIC Research Grant 2020 awarded to both authors by the International Association of Conference Interpreters.

REFERENCES

Angelelli, Claudia, and Brian James Baer. 2016. Researching Translation and Interpreting. London: Routledge.

Bartłomiejczyk, Magdalena. 2006. 'Strategies of Simultaneous Interpreting and Directionality.' Interpreting 8 (2): 149–174.

Bradley, James V. 1958. 'Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design.' Journal of the American Statistical Association 53 (282): 525–528. Braun, Sabine, Elena Davitti and Catherine Slater. 2020. ''It's like being in bubbles': affordances and challenges of virtual learning environments for collaborative learning in interpreter education'. The Interpreter and Translator Trainer 14(5): 1–20.

Chmiel, Agnieszka. 2008. 'Boothmates Forever? – On Teamwork in a Simultaneous Interpreting Booth.' Across Languages and Cultures 9 (2): 261–276.

Chmiel, Agnieszka. 2021a. 'Effects of Simultaneous Interpreting Experience and Training on Anticipation, as Measured by Word-Translation Latencies.' Interpreting 23 (1): 18–44.

Chmiel, Agnieszka. 2021b. 'Eye-Tracking Studies in Conference Interpreting.' In The Routledge Handbook of Conference Interpreting, edited by Michaela Albl-Mikasa and Elisabet Tiselius, 1st ed., 457–70. London: Routledge.

Chmiel, Agnieszka, Przemysław Janikowski, and Agnieszka Lijewska. 2020. 'Multimodal Processing in Simultaneous Interpreting with Text: Interpreters Focus More on the Visual than the Auditory Modality.' Target 32(1): 37–58.

Chmiel, Agnieszka, and Agnieszka Lijewska. 2019. 'Syntactic Processing in Sight Translation by Professional and Trainee Interpreters.' Target 31 (3): 378–397.

Chmiel, Agnieszka, and Iwona Mazur. 2013. 'Eye Tracking Sight Translation Performed by Trainee Interpreters.' In Tracks and Treks in Translation Studies, edited by Catherine Way, Sonia Vandepitte, Reine Meylaerts, and Magdalena Bartłomiejczyk, 189–205. Amsterdam: Benjamins.

Collard, Camille and Marta Buján Navarro. 2021. ESIT research project on remote simultaneous interpreting: first overview of results. Accessed 12 September 2022. https://www.researchgate.net/publication/350957508_First_overview_of_results

Cuetos, Fernando, María Glez-Nosti, Analía Barbón, and Marc Brysbaert. 2011. 'SUBTLEXESP: Spanish Word Frequencies Based on Film Subtitles.' Psicológica 32: 133–143.

Defrancq, Bart, and Claudio Fantinuoli. 2021. 'Automatic Speech Recognition in the Booth: Assessment of System Performance, Interpreters' Performances and Interactions in the Context of Numbers.' Target 33 (1): 73–102.

Díaz-Galaz, Stephanie, Presentación Padilla, and María Teresa Bajo. 2015. 'The Role of Advance Preparation in Simultaneous Interpreting.' Interpreting 17 (1): 1–25.

Doherty, Stephen, and Jan-Louis Kruger. 2018. 'The Development of Eye Tracking in Empirical Research on Subtitling and Captioning.' In Seeing into Screens: Eye Tracking and the Moving Image, edited by Tessa Dwyer, Claire Perkins, Sean Redmond, and Jodi Sita, 46–64. London: Bloomsbury.

Dragsted, Barbara, and Inge Gorm Hansen. 2009. 'Exploring Translation and Interpreting Hybrids. The Case of Sight Translation.' Meta 54 (3): 588–604.

Duchowski, Andrew T., Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. 'The Index of Pupillary

Activity: Measuring Cognitive Load Vis-à-Vis Task Difficulty with Pupil Oscillation.' In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–13. Montreal: ACM.

Environmental Implementation Review. Accessed 12 September 2022. https://environment.ec.europa.eu/law-and-governance/environmental-implementation-review_en

Fantinuoli, Claudio. 2019. 'The Technological Turn in Interpreting: The Challenges That Lie Ahead.' In Proceedings of the Conference Übersetzen und Dolmetschen 4.0.—Neue Wege im Digitalen Zeitalter, edited by Wolfram Baur and Felix Mayer, 334–354. Berlin: BDÜ Fachverlag.

Fantinuoli, Claudio, Giulia Marchesini, David Landan, and Lukas Horak. 2022. 'KUDO

Interpreter Assist: Automated Real-Time Support for Remote Interpretation'. In Proceedings of the Translating and the Computer 43 Conference, 68–77. Geneva: Editions Tradulex.

Gieshoff, Anne Catherine. 2018. The Impact of Audio-Visual Speech Input on Work-Load in Simultaneous Interpreting. Doctoral dissertation, Universität Mainz.

Gile, Daniel. 2009. Basic Concepts and Models for Interpreter and Translator Training. Amsterdam: Benjamins.

GT Booth. Accessed 12 September 2022. https://www.gtmeeting.com/solutions/gtbooth Gunning, Robert. 1969. 'The Fog Index after Twenty Years.' Journal of Business Communication, 6 (2): 3–13.

Hale, Sandra, Jane Goodman-Delahunty, Natalie Martschuk, and Stephen Doherty. 2022. 'The Effects of Mode on Interpreting Performance in a Simulated Police Interview.' Translation and Interpreting Studies 17 (2): 264–286.

Huang, Chih-Chieh. 2011. Tracking Eye Movements in Sight Translation – the Comprehension Process in Interpreting. Graduate Institute of Translation and Interpretation. Unpublished master's thesis, National Taiwan Normal University.

Halverson, Sandra L., and Ricardo Muñoz Martín. 2021. 'The Times, They Are a-Changin':

Multilingual Mediated Communication and Cognition.' In Multilingual Mediated Communication and Cognition, edited by Ricardo Muñoz Martín and Sandra L. Alverson, 1–17. London: Routledge.

Hart, Sandra G. and Lowell E. Staveland. 1988. 'Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research'. In Human Mental Workload, edited by Peter A. Hancock and Najmedin Meshkati, North-Holland, 139–183.

Hoberg, Felix. 2022. Informationsintegration in Mehrsprachigen Textchats. Der Skype Translator im Sprachenpaar Katalanisch-Deutsch. Berlin: Language Science Press.

Jakobsen, Arnt Lykke, and Kristian Jensen. 2008. 'Eye Movement Behaviour across Four Different Types of Reading Task.' In Looking at Eyes – Eye Tracking Studies of Reading and Translation Processing, edited by Susanne Göpferich, Arnt Lykke Jakobsen, and Inger Mees, 103–124. Copenhagen: Samfundslitteratur.

Jakobsen, Arnt Lykke. 2017. 'Translation Process Research.' In The Handbook of Translation and Cognition, edited by Aline Ferreira and John W. Schwieter, 19–49. Hoboken: Wiley and Sons.

Jones, Roderick. 2002. Conference Interpreting Explained. Manchester: St. Jerome.

Kajzer-Wietrzny, Marta, Ilmari Ivaska, and Adriano Ferraresi. 2021. ''Lost' in Interpreting and 'Found' in Translation: Using an Intermodal, Multidirectional Parallel Corpus to Investigate the Rendition of Numbers.' Perspectives: Studies in Translation Theory and Practice 29 (4): 469–88.

Korpal, Paweł. 2017. Linguistic and psychological indicators of stress in simultaneous interpreting. Poznań: Wydawnictwo Naukowe UAM.

Korpal, Paweł, and Katarzyna Stachowiak-Szymczak. 2018. 'The Whole Picture: Processing of Numbers and Their Context in Simultaneous Interpreting.' Poznan Studies in Contemporary Linguistics 54 (3): 335–54.

Korpal, Paweł, and Katarzyna Stachowiak-Szymczak. 2019. 'Combined Problem Triggers in Simultaneous Interpreting.' Perspectives: Studies in Translation Theory and Practice, 1–18.

Laugwitz, Bettina, Theo Held, and Martin Schrepp. 2008. 'Construction and Evaluation of a User Experience Questionnaire'. In HCI and Usability for Education and Work, edited by Andreas Holzinger, 5298:63–76. Berlin, Heidelberg: Springer Berlin Heidelberg.

Limesurvey GmbH (2003). / LimeSurvey: An Open Source survey tool /LimeSurvey GmbH, Hamburg, Germany. http://www.limesurvey.org. Accessed 28 September 2022.

Lin, Yumeng, Qianxi Lv, and Junying Liang. 2018. 'Predicting Fluency With Language Proficiency, Working Memory, and Directionality in Simultaneous Interpreting.' Frontiers in Psychology 9: 1543.

Liu, Minhua, Diane L. Schallert, and Patrick J. Carroll. 2004. 'Working Memory and Expertise in Simultaneous Interpreting.' Interpreting 6 (1): 19–42.

Liu, Yanmei, Binghan Zheng, and Hao Zhou. 2019. "Measuring the Difficulty of Text Translation: The Combination of Text-Focused and Translator-Oriented Approaches". Target: International Journal of Translation Studies 31 (1): 125–149.

Mellinger, Christopher D., and Thomas A. Hanson. 2017. Quantitative Research Methods in Translation and Interpreting Studies. Abingdon: Routledge.

Mellinger, Christopher D., and Thomas A. Hanson. 2018. 'Interpreter Traits and the

Relationship with Technology and Visibility.' Translation and Interpreting Studies. 13 (3), 366–392.

Mellinger, Christopher D., and Thomas A. Hanson. 2022. 'Considerations of Ecological Validity in Cognitive Translation and Interpreting Studies.' Translation, Cognition & Behavior, 1–26.

Mellinger, Christopher D. 2019. 'Computer-Assisted Interpreting Technologies and Interpreter Cognition: A Product and Process-Oriented Perspective'. Revista Tradumàtica. Tecnologies de la Traducció, 17, 33–44. < >

Moser-Mercer, Barbara. 2005. 'The crucial role of presence'. Bulletin VALS-ASLA 81: 73–97.

Mouzourakis, Panayotis. 2006. 'Remote Interpreting: A Technical Perspective on Recent Experiments'. Interpreting 8 (1): 45–66.

Naeeri, Salem, Saptarshi Mandal, and Ziho Kang. 2019. 'Analyzing Pilots' Fatigue for Prolonged Flight Missions: Multimodal Analysis Approach Using Vigilance Test and Eye Tracking.' In Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63 (1): 111–15.

NASA-TLX. 1986. NASA-TLX Paper and Pencil Package. https://humansystems.arc.nasa.gov /groups/tlx/downloads/TLX_pappen_manual.pdf. Accessed 28 September 2022.

Plevoets, Koen, and Bart Defrancq. 2016. 'The Effect of Informational Load on Disfluencies in Interpreting: A Corpus-Based Regression Analysis.' Translation and Interpreting Studies 11 (2): 202–24.

Prandi, Bianca. 2022. Computer-Assisted Simultaneous Interpreting. A Cognitive-Experimental Study on Terminology. Berlin: Language Science Press.

Ratchford, Mark and Michelle Barnhart. 2012. 'Development and Validation of the Technology Adoption Propensity (TAP) Index.' Journal of Business Research 65 (8): 1209–1215.

Rojo López, Ana María, Paula Cifuentes Férez and Laura Espín López. 2021a. '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.

Rojo López, Ana María, Ana Isabel Foulquié-Rubio, Laura Espín López and Francisco Martínez Sánchez. 2021b. 'Analysis of speech rhythm and heart rate as indicators of stress on student interpreters'. Perspectives: Studies in Translation Theory and Practice 29 (4): 591–607.

Roziner, Ilan, and Miriam Shlesinger. 2010. 'Much Ado about Something Remote: Stress and Performance in Remote Interpreting'. Interpreting 12 (2): 214–247.

Rozkrut, Paulina. 2022. Factors Influencing the Processing of Numbers in Polish-English and English-Polish Simultaneous Interpreting. Master's thesis, Adam Mickiewicz University in Poznań.

Saeed, Muhammad Ahmed, Eloy Rodríguez González, Tomasz Korybski, Elena Davitti, and Sabine Braun. 2022. 'Connected yet Distant: An Experimental Study into the Visual Needs of the Interpreter in Remote Simultaneous Interpreting'. In Human-Computer Interaction. User Experience and Behavior, edited by Masaaki Kurosu, 214–232. Cham: Springer International Publishing.

Saldanha, Gabriela, and Sharon O'Brien. 2014. Research Methodologies in Translation Studies. London: Routledge.

Schrepp, Martin. 2019. User Experience Questionnaire Handbook. https://www.ueq-online.org /Material/Handbook.pdf. Accessed 28 September 2022.

Seeber, Kilian G. 2012. 'Multimodal Input in Simultaneous Interpreting. An Eye-Tracking Experiment.' In Proceedings of the 1st International Conference TRANSLATA, Translation & Interpreting Research: Yesterday – Today – Tomorrow. Innsbruck, Austria, edited by

Lew N. Zybatov, Alena Petrova, and Michael Ustaszewski, 341–47. Frankfurt: Peter Lang.

Seeber, Kilian G. 2017. 'Interpreting at the European Institutions: Faster, Higher, Stronger.' CLINA 3 (2): 73–90.

Seeber, Kilian G., Laura Keller, Rhona Amos, and Sophie Hengl. 2019. 'Expectations vs. Experience: Attitudes towards Video Remote Conference Interpreting'. Interpreting, 21 (2): 270–304.

Seeber, Kilian G., Laura Keller, and Alexis Hervais-Adelman. 2020. 'When the Ear Leads the Eye – the Use of Text during Simultaneous Interpretation.' Language, Cognition and Neuroscience 35 (10): 1480–1494.

Seeber, Kilian G., and Dirk Kerzel. 2011. 'Cognitive Load in Simultaneous Interpreting: Model Meets Data.' International Journal of Bilingualism 16 (2): 228–42.

Setton, Robin, and Andrew Dawrant. 2016. Conference Interpreting – A Trainer's Guide. Amsterdam: Benjamins.

Seubert, Sabine. 2019. Visuelle Informationen beim Simultandolmetschen. Berlin: Frank & Timme.

Shao, Zhangminzi, and Mingjiong Chai. 2020. 'The Effect of Cognitive Load on Simultaneous Interpreting Performance: An Empirical Study at the Local Level.' Perspectives: Studies in Translation Theory and Practice 29 (5): 1–17.

Shreve, Gregory M., Isabel Lacruz, and Erik Angelone. 2010. 'Cognitive Effort, Syntactic Disruption, and Visual Interference in a Sight Translation Task.' In Translation and Cognition, edited by Gregory M. Shreve and Erik Angelone, 63–84. Amsterdam/Philadelphia: Benjamins.

Spielberger, Charles, Richard Gorsuch, Robert E. Lushene, Peter R. Vagg and

Gerard A. Jacobs. 1983. Manual for the State-Trait Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.

Stachowiak, Katarzyna. 2016. Eye Movements and Gestures as Correlates of Language Processing in Consecutive and Simultaneous Interpreting. Department of Translation Studies. Doctoral dissertation, Adam Mickiewicz University in Poznań.

Su, Wenchao, and Defeng Li. 2019. 'Identifying Translation Problems in English-Chinese Sight Translation.' Translation and Interpreting Studies 14 (1): 110–34.

Sun, Sanjun and Gregory M. Shreve. 2014. 'Measuring translation difficulty: an empirical study'. Target 26 (1): 98–127.

Szarkowska, Agnieszka, and Olivia Gerber-Morón. 2019. 'Two or Three Lines: A MixedMethods Study on Subtitle Processing and Preferences.' Perspectives: Studies in Translation Theory and Practice 27 (1): 144–64.

Tiselius, Elisabet. 2009. 'Revisiting Carroll's Scales.' In Testing and Assessment in Translation and Interpreting Studies, edited by Claudia V. Angelelli and Holly E. Jacobson, XIV: 95–121. Amsterdam: John Benjamins Publishing Company.

Tiselius, Elisabet, and Kayle Sneed. 2020. 'Gaze and Eye Movement in Dialogue Interpreting: An Eye-Tracking Study.' Bilingualism: Language and Cognition 23 (4): 780–87.

Van Heuven, Walter J.B., Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. 'SUBTLEX-UK: A New and Improved Word Frequency Database for British English.' The Quarterly Journal of Experimental Psychology 67 (6): 1176–90.

Yang, Chengsong, Yimeng Wang and Na Fan. 2022. “Are parallel translation tasks parallel in difficulty? an eye-tracking study”. Perspectives 30 (4): 711–726.