QUANTIFYING THE IMPACT OF SPEECH INTELLIGIBILITY ON TASK PERFORMANCE VIA PHYSIOLOGICAL SIGNALS

A Thesis

by

SHRAVANI SRIDHAR

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Theodora Chaspari
Committee Members,	James Caverlee
	Winfred Arthur, Jr.
Head of Department,	Scott Schaefer

May 2022

Major Subject: Computer Science

ABSTRACT

Background noise can potentially cause dissatisfaction, stress, and reduced performance in the workplace. One type of background noise comes from background discussions. Many studies in the literature exist on the effects of speech intelligibility, a common measure of speech comprehension. Presently, no unitary framework exists for describing the relationship between stress and performance; furthermore, the exact constitution of stress response in terms of physiological and subjectively measured characteristics is not yet determined.

This thesis investigates the impact of various speech intelligibility conditions on cognitive task performance in a simulated office environment using electrodermal activity (EDA) and heart rate (HR) physiological signals. It was hypothesized that increase in distraction causes increase in stress, which causes decrease in performance, and thus distraction and performance are inversely related. A user study was conducted in a simulated office setting where participants (N = 24, N = 29) performed different cognitive tasks listening to audio of variable speech intelligibility conditions. Data collected from participants included EDA and HR signals and self-reported subjective ratings during the experiment, and task performance scores. Correlations were computed to help assess conformance of the data to the hypothesized relationships. Physiological signals and self-reported subjective ratings were hypothesized as indicators of stress; correlations were computed to help determine how stress could be measured between them. This thesis also aims to determine which intelligibility conditions have a stronger effect on stress and performance, and how certain findings are affected by the type of cognitive task and by experimental design variations, namely, (1) Within-Subjects design and (2) Between-Groups (randomized) design.

The following were found from the analysis results. Distraction was not directly related to stress in either experimental design. Stress was inversely related to performance in both designs. Distraction was inversely related to performance in the second design. Physiological signals and self-reported stress ratings were a more accurate stress indicator in the first and second design, respectively. The high intelligibility condition overall affected performance strongest in the second design but not stress in either design. Overall the Between-Groups (randomized) design yielded results more in line with our research hypotheses than the Within-Subjects design.

DEDICATION

To my mother, father, and sister.

ACKNOWLEDGMENTS

I am extremely grateful to my advisor, Dr. Theodora Chaspari, for providing me the opportunity to pursue this thesis under her guidance. I am deeply indebted to her for her endless support, help, and patience throughout my process of conducting the research and writing the thesis. Her research knowledge and expertise will always serve as inspiration to me.

I am immensely grateful to my committee member, Dr. Winfred Arthur, Jr., for his indispensable involvement in my thesis study; his invaluable suggestions and advice, as well as vast knowledge and experience in his field, helped steer me towards the completion of my thesis.

I would like to express my deepest appreciation for my committee member, Dr. James Caverlee, for his greatly insightful and thought-provoking feedback that helped enliven the work and implications of this thesis and that also encourages me to raise my thesis to a higher level.

I would like to convey special thanks to Ehsanul Haque Nirjhar, my fellow researcher at HUman Bio-Behaviour Signals (HUBBS) Lab, for his valuable and always prompt assistance during my thesis study. His knowledge and experience served as a foundation for many crucial aspects of my thesis. I would also like to specially thank Evan Nault, Ellen Hagen, and Gracie Cooper for taking time out of their schedules to help me out when I had questions regarding the user study and data collection; their assistance was immensely significant for my research. I would like to thank all of the research team for their extremely hard work in putting together the user study, and for providing me with all relevant data needed to conduct my research analysis; I am deeply grateful for your help.

I have learned so much from each and everyone mentioned. Working on this thesis with their help helped to expand my mind and teach me to always remain humble and grateful.

I would like to thank Texas A&M University for providing me an opportunity to pursue a Master of Science degree in Computer Science.

Finally, I would like to thank my friends and my family for their constant support and encouragement throughout the research, which helped fuel and motivate me to complete my thesis, as

well as helped serve as a reminder of my love of research during the tough moments. My heartfelt gratitude to all.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

The thesis committee for this work includes Dr. Theodora Chaspari (Chair), Dr. James Caverlee of the Department of Computer Science and Engineering at Texas A&M University, and Dr. Winfred Arthur, Jr. of the Department of Psychological and Brain Sciences at Texas A&M University. The user study for this research was conducted with the help of members of the research team, including Dr. Theodora Chaspari, Dr. Winfred Arthur, Jr., Ellen Hagen, Evan Nault, Gracie Cooper, and Benjamin Schulte. The data used in the analysis in this research was provided and finalized with the help of Dr. Theodora Chaspari, Dr. Winfred Arthur, Jr., Ellen Hagen, Evan Nault, Gracie Cooper, and Benjamin Schulte. Base code files needed for EDA and HR feature extraction were provided by Ehsanul Haque Nirjhar. All other work conducted as part of this thesis was completed by the author independently.

Funding Sources

The user study for this research has been supported by the School of Innovation at Texas A&M University through the Innovation[X] program. Any opinions, findings, conclusions, and recommendations expressed in this thesis are those of the author and do not necessarily represent those of the School of Innovation.

NOMENCLATURE

EDA	Electrodermal Activity
HR	Heart Rate
SCR	Skin Conductance Response
ECG	Electrocardiogram
GSR	Galvanic Skin Response
STI	Speech Transmission Index
SNS	Sympathetic Nervous System
PNS	Parasympathetic Nervous System
ANS	Autonomic Nervous System
HF-HRV	High-Frequency Heart Rate Variability
SWELL-KW	Smart Reasoning Systems for Well-being at Work and at Home - Knowledge Work dataset
GMA	General Mental Ability
STAI	State-Trait Anxiety Inventory
BVP	Blood Volume Pulse
NASA-TLX	NASA Task Load Index

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Exposure to prolonged noise can have a detrimental impact on productivity and cause a variety of health problems, including stress and cardiovascular diseases [1]. Background noise is a common stressor and distractor, and it negatively affects tasks that involve cognition, attention, and concentration, all of which are necessary to perform well in the workplace [2], [3], [4]. Previous studies suggest that workers who are dissatisfied with the office environment are less productive [5], [6], [7], [8], [9], [10], [11]. In addition, the more control people feel they have over their environmental conditions, the more comfort and satisfaction they feel which leads to increased perceived productivity [12]. Perceived noise annoyance has been shown to increase arousal levels [13]. Noise in the work environment can induce the release of cortisol resulting in increased arousal that can hinder reasoning and attention [14], [15]. In fact, noise distraction even at realistic levels in an open-plan office setting leads to increase in fatigue with negative effects on the work performance [16]. All these results show that, not just background noise, but even perceived noise, has a significant negative effect on performance. However various noise types do not affect human performance in the same manner: noise energy concentrated in low frequencies may hinder verbal reasoning tasks [10], increased intelligibility of noise coming from speech can be more detrimental to human performance [17], [18], while excessive noise absorption in open-plan office environments might cause reduced acceptability to sporadic noises [19].

Background speech noise in particular can negatively impact speech processing ability [20], while speech has been shown to negatively impact reading and short-term memory performance [21]. Speech has also been perceived as the most distracting source of noise in office environments, particularly in open-plan office settings [18], [22]. Further, high intelligibility background speech negatively impacts performance on the basic cognitive tasks of verbal-logical reasoning, verbal short-term memory, and sustained attention, more than lower intelligibility levels and silence [23]. In addition, it has been shown that increasing background speech intelligibility results in performance decline in many cognitive tasks [24]. Reducing background noise in open-plan offices can further alleviate the negative effects of psychological job stress [25]. These results indicate the importance of reducing the negative effects of background speech noise, especially in open-plan office environments.

Furthermore, the effect of environmental noise on productivity can be different across individuals: the perceived sensation differs from person to person and therefore a single environmental condition cannot have the same effect on all people [26]. These findings from psychological sciences demonstrate the need for data-driven personalized approaches that can help identify the levels of noise that can be tolerated by each person, which can contribute to customized solutions and adaptive environments for mitigating noise. These are further amplified in the context of an open-plan office environment due to high variability in the types of environmental stimuli and individual preferences.

The goal of this Master's thesis is to investigate how various conditions of speech intelligibility of background discussions affect stress and work performance. A user study was conducted in a simulated office setting where participants performed different cognitive tasks while simultaneously listening to audio simulating background discussions of variable conditions of speech intelligibility. Various physiological signals were collected from the participants using sensors during the tasks, and their various self-reported subjective ratings were obtained via self-reported measures at various times of the experiment. The participants' task performance scores were also obtained. The speech intelligibility conditions employed in the study are, namely, high intelligibility, low intelligibility, and silence conditions. The participants' obtained physiological signals, self-reported measures, and task performance measures were analyzed to understand the impact of these background distractors on their performance and stress while performing the cognitive tasks.

To investigate the above, this thesis hypothesizes a set of relationships among distraction, stress, and performance and examines these on data from the user study. (It is important to mention here that distraction and performance are assumed in this research to refer to speech distraction and work performance, respectively, and are thus used as so henceforth in this thesis.) Based on aforementioned prior research, and given that research has shown that increase in stress can lead to decreased employee performance [27], it is hypothesized that increase in distraction causes increase in stress, which in turn causes decrease in performance, and that thus distraction and performance are inversely related. Assuming that distraction, represented by the speech intelligibility of the background discussions, is X, and its outcome is the work performance under the various background discussions, Y, represented by the participants' performance measures on the cognitive tasks under the background distractors, and stress, represented by the participants' self-reported subjective ratings and their collected physiological signals during the tasks, is M, then we make the following hypotheses:

1. Increase in distraction X causes increase in stress M.

2. Increase in stress M causes decrease in performance Y.

3. Distraction X and performance Y are thus inversely related.

In addition, it has been theorized in the literature that stress response constitutes physiological characteristics, and also that one aspect of stress includes subjective responses to the stress [28]; based on these, it is hypothesized in this thesis that physiological signals and self-reported subjective ratings are two indicators of stress.

In my Master's thesis, I aim to investigate whether the above hypotheses are reflected in data collected from the aforementioned user study. Additionally, I aim to investigate how certain results of the corresponding analyses are affected by the type of cognitive task as well as by experimental design variations. Furthermore, I also aim to determine which intelligibility conditions have a stronger effect on stress and performance.

Thus, in this thesis I aim to answer the following research questions:

1. Is distraction related to stress?

2. Is stress related to performance?

3. Is distraction related to performance?

4. Does it matter how stress is operationalized or measured between the self-reported subjective ratings and the physiological signals?

5. Does the specific type of task make a difference in any of the above results?

6. Which intelligibility conditions have a stronger effect on stress and performance?

7. Are these results consistent across different design variations of the experiment?

2. LITERATURE REVIEW AND RESEARCH CONTRIBUTIONS

2.1 Stress and relationship between stress and performance

There have been many theories in the literature aiming to describe how performance is affected by stress, but there is no unitary framework that has consensus in the scientific community for the same as of yet. One such theory is the Inverted-U hypothesis [28]. The Inverted-U hypothesis describes performance as increasing with arousal, until a point after which performance decreases; the hypothesis emerged from the postulate that the relationship between performance and arousal is curvilinear in the shape of an inverted U, representing that moderate arousal leads to optimal performance, while too much or too little arousal will lower performance. But later observations, including one that the Inverted-U hypothesis does not explain how the curvilinear inverted U relationship is produced, as well as another that the hypothesis is based on seemingly psychologically trivial research, to name a few, refuted this theory, leading to newer theories. Many of these latter theories are supported by the literature. One of these theories suggests that performance is dependent on the type of stress stimuli and how the performance is being measured. There is empirical evidence for other models that likewise emerged later, attempting to describe performance under stress, including those representing non-inverted U relationships, linear relationships (positive as well as negative), and no or nearly no relationship at all. As stated earlier, there is no consensus as of yet from the scientific community on a single unitary framework for describing how stress impacts performance; there are currently only theories being proposed and disputed.

There is also no unitary definition for stress in the scientific community as of yet, including its constitution [28]. However, there are many theories in the literature regarding the same. Arousal theory states that arousal activates the stress response, which is usually a multidimensional response that frequently includes physiological dimensions [28]. Wofford *et al.* defined stress response in humans as comprising three domains, namely, physiological arousal, psychological responses, and behavioral responses [29]. Gaillard stated that stress is multi-dimensional, with one aspect of stress being an output function that includes subjective and physiological responses to the stress [30]. Examples of physiological responses to stress are HR, skin conductance response (SCR), or EDA, and electrocardiogram (ECG), while examples of subjective responses to stress are comfort, cognitive demand, and frustration [28].

Relevance to this research: In the work of my thesis, I am not attempting to suggest or prove any one of the aforementioned theories in the literature as being the unitary or superior one for describing the relationship between stress and performance or for defining the constitution of stress, nor am I attempting to suggest or prove a novel theory for these purposes; that is not the goal of this thesis. However, some of these theories could help explain the potential patterns that might emerge from the data analysis conducted for answering the research questions of this thesis. In this research, as previously mentioned, it is hypothesized that the relationship between stress and performance is positive linear, and that stress is indicated by physiological signals and self-reported subjective ratings; these hypotheses are subsequently examined on the data for the purpose of answering the research questions.

2.2 Relationship between physiological signals, and stress and subjective ratings

Bong et al. proposed a system to assess and classify the emotional stress levels of individuals using their physiological signals, including ECG, and their perceived stress ratings [31]. Studies have shown that stress and fatigue can be reliably predicted from physiological signals [32]. One study has found that the physiological signals of ECG and galvanic skin response (GSR), or SCR, are two of the more accurate predictors of an individual's stress and fatigue while performing real-world driving tasks [32]. Another study showed the potential usefulness of real-time ECG measurements for detecting sudden rise and high levels of emotional response, mental overload, and physical activity for workers in professions that involved risks and/or responsibility for people's lives [33]. Heart rate, along with anxiety, are increased in response to acute psychological stress [34]. Subjective workload is positively correlated with anxiety, which is in turn positively correlated with heart rate [35]. Furthermore, a study in real flight showed that heart rate was significantly positively correlated with mental workload and stress in the pilots; mental workload and stress were found to be positively correlated with each other as well [36]. Brookhuis *et al.* showed that, with the help of heart rate measures from ECG, obtaining drivers' mental workload measurements in driving simulators is relatively feasible using physiological signals [37]. Brookhuis *et al.* also stated that physiological measures are the most natural indicator for mental workload [37]. Mehler *et al.* showed that increase in task demands during simulated driving caused increases in EDA, HR, and respiration rate [38]. Additionally, EDA was found to distinguish between cognitive, or mental, workload and stress in a simulated office setting study with reasonably high accuracy [39].

Exposure to high-intensity aperiodic noise can lead to increase in electrodermal activity and decreased task performance when workers believe stopping the noise is not in their control, than when they do believe that stopping the noise is in their control under the exact same setup conditions [40]. Previous research has also explored the effect of short-term acoustic stimuli on human physiology, indicating increased electrodermal activity and decreased heart rate under steady-state noise [41], [42]. Recent studies have examined this association in real-life residential building noises, indicating that individuals with higher noise sensitivity are physiologically more reactive and depict slower habituation to noisy conditions compared to the low sensitivity group [42], [43], [44]. Such associations are moderated by age, sex, and personality [45].

Relevance to this research: This thesis analyzes the impact of distractors on stress and cognitive task performance using EDA and HR data and self-reported subjective ratings. As mentioned previously, stress is hypothesized in this research to be indicated by physiological signals and self-reported subjective ratings under the distractor conditions, and this hypothesis is subsequently examined on the data.

2.3 Effects of speech intelligibility on performance

One study reported that, among intelligibility conditions of background speech noise, where the speech intelligibility was measured using Speech Transmission Index (STI), varying from STI = 0.0 to STI = 1.0, the greatest decline in performance occurred under the perfectly intelligible speech condition (STI = 1.0), while performance was the best when speech was absent from background noise (STI = 0.0) [20]. Additionally, increasing background speech intelligibility results in performance decline in many cognitive tasks [23]. Background speech has been shown to have harmful effects on short-term memory and working memory cognitive performance [23]. Hongisto *et al.* stated that the distracting power of speech, in terms of effect on performance, is determined by the speech intelligibility, which can be measured using STI [20]. Jahncke *et al.* showed that the relationship between STI and performance, as well as the amount of change in performance, in an open-plan office, are dependent on the type of cognitive task being considered [46]. Further, the results of the aforementioned study by Jahncke *et al.* showed that the magnitude of increase in cognitive performance from minimizing speech intelligibility would vary depending on the task type [46].

Relevance to this research: This work aims to analyze the impact of different speech intelligibility conditions on stress and cognitive task performance, and also at the task-type level. Based on the aforementioned prior work, it is premised in this research that greater intelligibility conditions result in greater distraction, and the subsequent analysis in this thesis as well as answering the research questions are based on this premise.

2.4 Assessing speech using physiological signals

Physiological stress responses such as the activation of the sympathetic nervous system (SNS) and the disengagement of the parasympathetic nervous system (PNS) may be evoked by cognitive and emotional challenges [47]. These physiological changes occur in the autonomic nervous system (ANS), which consists of the SNS and the PNS [47]. The SNS is associated with preparing the body for fight or flight in response to stress, while the PNS tries to restore the body to a calm state [47]. According to Mackersie *et al.*, EDA is known to reflect sympathetic activity, while high-frequency heart rate variability (HF-HRV) is known to reflect parasympathetic activity [47]. A study by Mackersie *et al.* showed that an increase in speech rate of background speech during auditory tasks, which results in an increase in auditory task demand, led to an increase in skin conductance levels (signifying the arousal of the SNS) and a decrease in HF-HRV (signifying the disengagement of the PNS) [47].

A study by Francis *et al.* reported that certain masked speech conditions instituted higher task demands and were significantly less intelligible than unmasked natural speech settings [48]. In the aforementioned study by Francis *et al.*, although the subjective task demand ratings and performance measures were comparable across two masked speech conditions differing in how the maskers were created, SCR was significantly greater in the condition where the masker had higher speech intelligibility [48].

Relevance to this research: This thesis analyzes cognitive task performance under different speech intelligibility conditions of simulated background discussions. It also hypothesizes that physiological signals are an indicator of stress response under the different speech intelligibility conditions and examines how the data reflects this.

Research contributions: This Master's thesis will examine physiological responses to noisy stimuli in association with task performance. The research design is modeled after prior work that has attempted to quantify task performance from physiological data (i.e., SWELL-KW dataset [49]), including tasks similar to those in this Master's thesis study, such as preparing reports and presentations, searching for information, and reading emails. Instead of audio distractors, prior work [49] has used stressors of time pressure and email interruptions for providing distraction. Furthermore, to the best of my knowledge, the work as part of my thesis is the first in the literature to use physiological signals as indicators of stress to analyze impact of background speech intelligibility on cognitive task performance.

3. USER STUDY

3.1 User study structure

The data analyzed in this thesis are part of a user study that was conducted in a research laboratory in the Department of Psychological & Brain Sciences at Texas A&M University. Participants with the following eligibility requirements were recruited: (1) being at least 18 years of age; (2) being a Native English speaker; and (3) having normal hearing capabilities. A controlled lab setting which simulated an office environment was set up, where participants spent approximately 3 hours performing different cognitive tasks in front of a desktop computer. The tasks included the following: (1) Proofreading/Catalog task (comprising of Proofreading and Catalog sub-tasks), (2) Presentation task, and (3) Prioritizing Task. During the tasks, the participants listened to various audio provided for them with correspondingly assigned speech intelligibility conditions, and various physiological data were recorded using sensors. In addition to guidelines regarding the requirements for the reports and presentations, they were given Internet access and auxiliary resources for completing the tasks. Participants were told in advance of signing up for the study that they would be given a $30 Amazon gift card for their participation in the research.

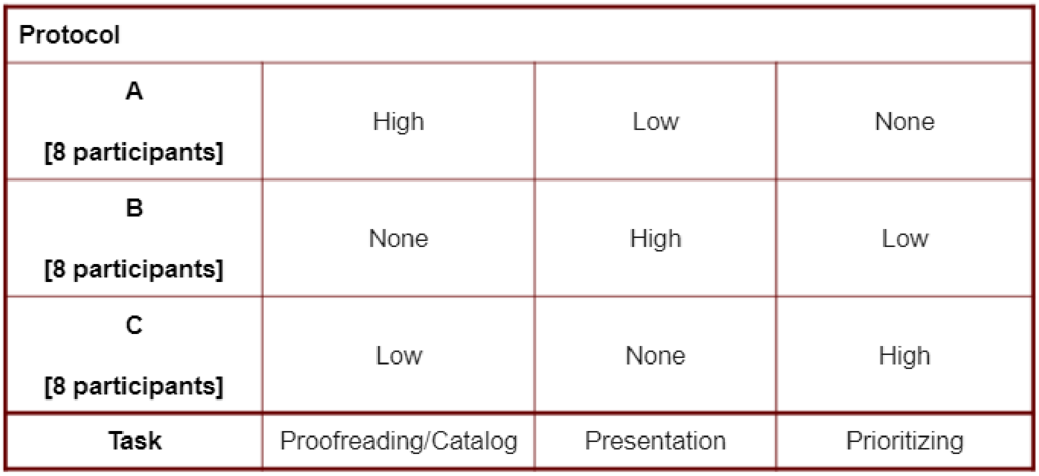
3.1.1 Speech intelligibility conditions

As mentioned previously, the speech intelligibility conditions used in the study are high intelligibility, low intelligibility, and silence conditions. Audio files to be used for simulating background discussions corresponding to the different speech intelligibility conditions were chosen by the research team. The process in which this was done involved first carefully listening to various candidate audio files for each intelligibility condition, and then eliminating any files with content that was not neutral and unbiased, both in terms of information conveyed and emotion, as well as files wherein a speaker had a discernible accent. The purpose of this elimination was to avoid any audio that would cause confounding distraction and/or reactions in the participants. The team then chose files for each speech intelligibility condition from the remaining candidates based on a consensus on perceived intelligibility. The STIs of the candidate audio files were also calculated; these STI values were inspected on the basis of how reasonable they were for the assigned intelligibility condition, the results of which were taken into consideration as well in selecting the final audio files for each intelligibility condition. The silence condition was intended to contain no background discussions (as well as no background noise) and be of zero intelligibility; hence, in order to simulate this condition, no audio files were selected for it. The final set of assigned audio files were of variable STIs corresponding to the high and low intelligibility conditions, respectively, and it was these files that were finally employed in the study. The STIs of these final files used for each intelligibility condition fell within different ranges corresponding to the condition. Specifically, the STIs of the audio files used for the high intelligibility condition were in the approximate range of STI = 0.24 to STI = 0.27, while the STIs of those used for the low intelligibility condition were in the approximate range of STI = 0.2 to STI = 0.22. The STI of the silence condition is taken to be 0.0.

3.1.2 Experimental design variations

The experiment protocol was divided into three blocks, where each block was assigned one cognitive task and one speech intelligibility condition. The experiment was conducted with two variations in this design, with different participants recruited for each: (1) Within-Subjects design, where all the participants performed each task under each speech intelligibility condition, and the task order was the same for all the participants, but the order of the intelligibility conditions was varied among them, and (2) Between-Groups (randomized) design, where each group of participants performed all the tasks under only one specific intelligibility condition, and the task order was varied randomly within the assigned groups. In the first experimental design, exactly one-third of the total number of participants were assigned in each order of speech intelligibility conditions. In both experimental designs, the Proofreading/Catalog task was 22 minutes long, whereas the Presentation and Prioritizing tasks were 20 minutes long each. Different participants were recruited for each of the design variations, namely, 24 subjects participated in the first design of the experiment, while the second design of the experiment had 29 participants. The second experimental design was conducted after the completion of the first experimental design.

Figures 3.1 and 3.2 show the setups followed for participants in the first and second experimental design variations, respectively, in tabular form.



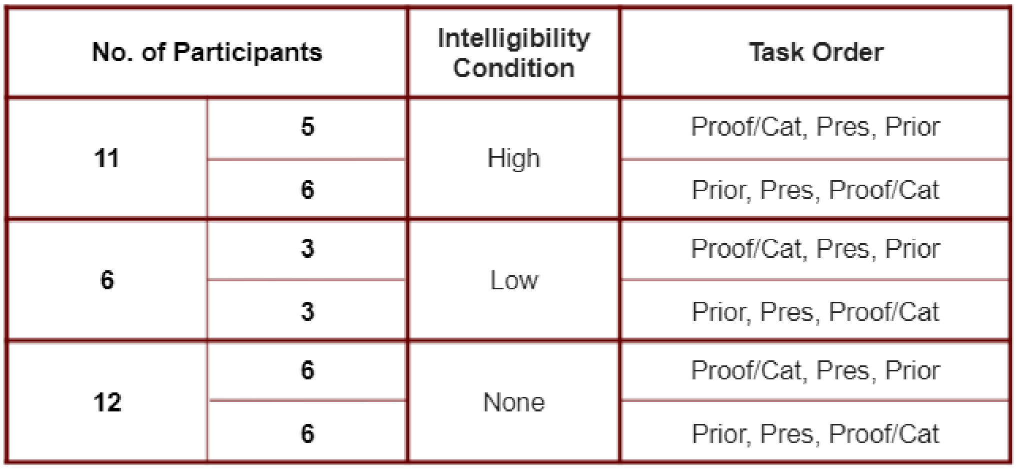
Protocol			
A [8 participants]	High	Low	None
B [8 participants]	None	High	Low
C [8 participants]	Low	None	High
Task	Proofreading/Catalog	Presentation	Prioritizing

Figure 3.1: The setup followed in the first experimental design of the user study (24 total participants).

In the below subsections, we describe the procedure of the experiment followed for both experimental design variations.

3.2 Pre-study

At the start of the experiment, participants were made to go through Covid-19 screening and review and sign a Covid-19 screening consent form. After this they were told to review and sign an Informed Consent form regarding the study. In order to test whether their hearing capabilities satisfied the eligibility requirements, they were made to take a hearing test, after which only those who passed the test were allowed to stay for the experiment. Subsequently, general mental ability (GMA) was captured via a cognitive assessment. They further filled out a set of measures to capture other individual differences measures, as listed below:



No. of Participants		Intelligibility Condition	Task Order
11	5	High	Proof/Cat, Pres, Prior
	6		Prior, Pres, Proof/Cat
6	3	Low	Proof/Cat, Pres, Prior
	3		Prior, Pres, Proof/Cat
12	6	None	Proof/Cat, Pres, Prior
	6		Prior, Pres, Proof/Cat

Figure 3.2: The setup followed in the second experimental design of the user study (29 total participants); the abbreviations of the task names in the figure, namely, Proof/Cat, Pres, and Prior, correspond to the tasks of Proofreading/Catalog, Presentation, and Prioritization, respectively.

- Trait form of the State-Trait Anxiety Inventory (STAI-Trait) [50], [51]: This is a 20-item measure for assessing trait anxiety, rated on a 4-point scale. Higher scores represent higher levels of anxiety. Examples of questions are "I feel rested" and "I feel inadequate".

- Five-Factor Model of Personality [52], [53]: This assessment measures the "Big Five" personality dimensions, namely, Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience.

- Coping Inventory for Stressful Situations Short Form [54], [55]: This measure assesses a person's coping in terms of three coping methods: task-oriented coping, emotion-oriented coping, and avoidance coping. Task-oriented coping is coping through applying task-oriented actions for the purpose of solving the task or problem. Emotion-oriented coping is coping through self-oriented emotional responses aimed to alleviate stress. Avoidance coping is coping through distraction from the task in order to avoid or reduce the stress.

- Weinstein's Noise Sensitivity Scale [56]: This is a measure for assessing a person's sensitivity to noise, rated on a 6-point scale. Examples of questions are "I am easily awakened by noise." and "There are often times when I want complete silence."

3.2.1 Wearable sensors

After filling out the above individual differences measures, participants were equipped with the Empatica E4 [57] and the Actiwave Cardio [58] sensors, which were used for recording participants' various physiological signals during the performance tasks. These devices are described below, along with measures they recorded in the user study:

- Empatica E4 wristband: This sensor allows various physiological data to be obtained in real-time and is worn on the wrist [57]. Participants' EDA, HR, and blood volume pulse (BVP) signals were recorded using this sensor during the cognitive tasks.

- Actiwave Cardio: This is a chest sensor and was used for recording participants' ECG waveforms during the cognitive tasks [58].

3.3 In-study

Next the participants followed the respective protocol assigned to them.

1. At the beginning of each block, they watched a 5-minute psychological relaxation video and then were asked to fill out the State form of STAI (STAI-State).

2. After that, they proceeded with completing the cognitive task, during which they performed the assigned task while listening to audio of the assigned intelligibility condition for the specific block and were also asked to rate their current stress levels via a pop-up single-item question every 5 mins.

3. After the task, participants also provided subjective ratings of stress, discomfort, and cognitive workload via STAI-State [50], a single-item discomfort scale, and the NASA Task Load Index (NASA-TLX) [59], in that order, respectively.

Participants were given a 5-minute break between blocks. Performance measures on the cognitive tasks were also obtained after the end of the experiment.

3.3.1 In-study assessments

The assessments filled out by the participants during the study are described below:

- Pop-up stress rating: This is a single-item measure developed for assessing the participants' subjective stress rating during the cognitive tasks in 5-minute intervals.
- State form of the State-Trait Anxiety Inventory (STAI-State) [50], [51]: This is a 20-item measure for assessing state anxiety, rated on a 4-point scale. Higher scores represent higher levels of anxiety. Examples of questions are "I feel strained" and "I am jittery".
- Discomfort: This is a single-item measure for assessing the subjective discomfort rating.
- NASA Task Load Index (NASA-TLX) [59]: This measure assesses 6 dimensions of workload, namely, Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration.

3.3.2 Cognitive tasks

Here we describe the three cognitive tasks used in the experiment, as well as corresponding evaluation and task performance measures relevant to the data analysis of the thesis:

- Proofreading/Catalog: This task comprised of two sub-tasks as follows:
  - In the Proofreading sub-task, participants were required to correctly proofread three cover letters. The errors were geared more towards aspects like grammar, spelling, sentence structure, and punctuation than towards form or structure of the entire document.

    For each cover letter, a point was added for each error corrected correctly, and a point was deducted for each error unaddressed or improperly corrected as well as for each new error that was made. The total score for this task was the sum of the scores of each of the three cover letters.

    - In the Catalog sub-task, participants were given four orders, as well as a product catalog containing all available items, their order numbers, and their prices. They were then required to correctly process the orders by correctly calculating each of their prices. Each order was given one point if its price was correctly calculated and 0 if not. The total score for this task was the sum of the scores of each of the four orders.

- Presentation: In this task, participants were required to make a brief oral presentation on the history and significance of Gobekli Tepe by preparing a Microsoft PowerPoint Presentation and a short script using Microsoft Word, which they were then required to present at the end of the session. Copy-and-paste was not allowed in the task.

  The presentation slides and the Word script were jointly given a basic criteria score as well as a quality score according to how well they matched certain preset basic and quality criteria, respectively. The total score for this task was the sum of the basic criteria score and the quality score.

- Prioritizing: In this task, participants were provided a set of tasks to prioritize for a hypothetical day, given that they had just arrived for work at 8 am on that particular day. They were required to list the tasks in the order that they would complete them on that day, and list tasks they would not have time for as those that would be done the next day (in no particular order).

  A predefined key was used to score the prioritization ranks both in terms of match to the key (scoring in absolute terms) and in terms of distance from the key (scoring in relative terms). The total absolute and relative scores of the task were the sum of prioritization rank absolute and relative scores, respectively.

The procedure of the in-study experiment, which was common to both experimental design variations, is summarized as a flowchart in Figure 3.3.

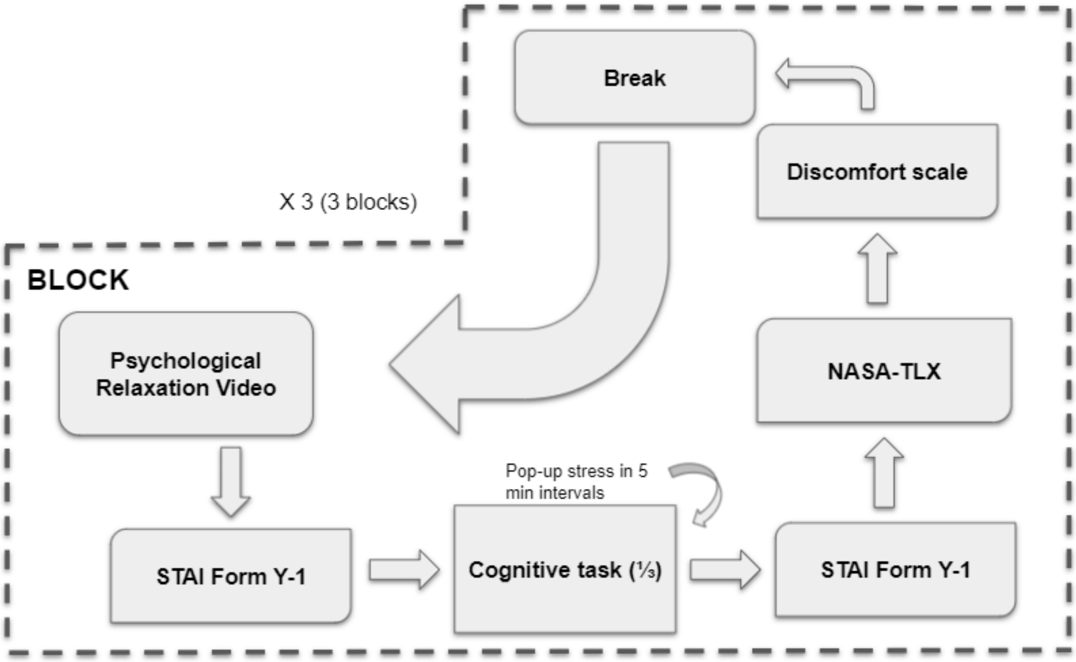


Figure 3.3: Flowchart describing the procedure of the experiment followed during in-study for both experimental designs.

4. METHODOLOGY

4.1 Data description

The data used for computing the analysis results of this Master's thesis is outlined as follows:

1. Physiological signals:
   - EDA: This is a measure of the changes in the electrical conductance of skin as a response to sweat secretion, recorded in microSiemens. Experiencing emotional arousal results in increase in skin conductance levels.
   - HR: This is a measure of the heart rate, recorded in beats per minute.
2. Self-reported measures (as described previously):
   - State Anxiety (from STAI-State)
   - Trait Anxiety (from STAI-Trait)
   - Pop-up stress
   - Discomfort
   - Cognitive workload from NASA-TLX
3. Task performance measures:
   - Proofreading total score: This is the total score for the Proofreading sub-task. The maximum possible values of the total scores for each of the three cover letters are 16, 16, and 14, respectively. Thus, based on this and the aforementioned evaluation procedure for the Proofreading sub-task, the maximum possible value of the total score is 46, the minimum possible value is negative infinity, and the possible values are separated by a step size of 1.

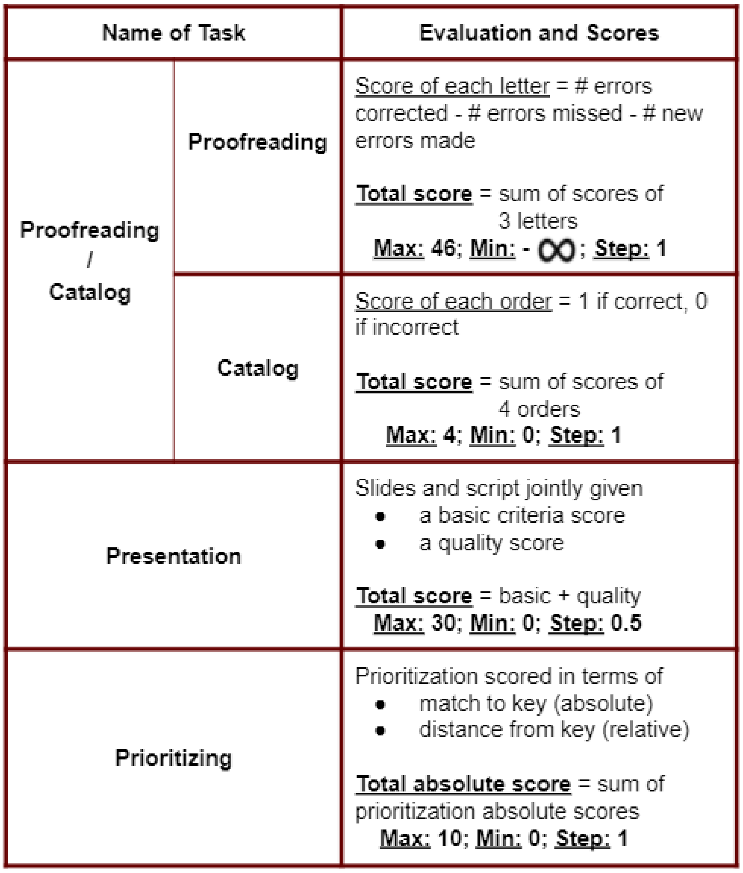
- Catalog total score: This is the total score for the Catalog sub-task and represents the total number of orders priced correctly out of the total four. Based on the aforementioned evaluation procedure for the Catalog sub-task, the maximum possible value of this score is 4, the minimum is 0, and the possible values are separated by a step size of 1.
- Presentation total score: This is the total score for the Presentation task. The basic criteria score was rated across 6 dimensions on a 3-point scale (0 - not met, 1 - partially met, and 2 - fully met) for a minimum possible of 0 and a maximum possible of 18, and the latter score was rated across 6 dimensions on a 5-point scale (0 - basic, 0.5, 1 - proficient, 1.5, 2 - advanced) for a minimum possible of 0 and a maximum possible of 12. Thus, based on this and the aforementioned evaluation procedure for the Presentation task, the possible values for the total score ranges from 0 to 30, separated by a step size of 0.5.
- Prioritizing total absolute score: This is the total absolute score for the Prioritizing task. Based on the aforementioned evaluation procedure for the Prioritizing task, the maximum possible value of this score is 10, the minimum is 0, and the possible values are separated by a step size of 1.

Figure 4.1 summarizes the relevant evaluation procedure and relevant score calculations and ranges for the tasks.

4.2 Physiological signal processing

4.2.1 Signal segmentation

The EDA and HR signals were first segmented per task, separately for each experimental design.



Name of Task		Evaluation and Scores
Proofreading / Catalog	**Proofreading**	Score of each letter = # errors corrected - # errors missed - # new errors made **Total score** = sum of scores of 3 letters **Max: 46; Min: - ∞; Step: 1**
	Catalog	Score of each order = 1 if correct, 0 if incorrect **Total score** = sum of scores of 4 orders **Max: 4; Min: 0; Step: 1**
Presentation		Slides and script jointly given • a basic criteria score • a quality score **Total score** = basic + quality **Max: 30; Min: 0; Step: 0.5**
Prioritizing		Prioritization scored in terms of • match to key (absolute) • distance from key (relative) **Total absolute score** = sum of prioritization absolute scores **Max: 10; Min: 0; Step: 1**

Figure 4.1: Summary of relevant evaluation procedure and relevant score calculations and ranges for each cognitive task/sub-task.

4.2.2 Signal pre-processing

The segmented signals were then inspected to make sure that unusable signals (for example, with abnormally low levels) were removed from the data prior to processing. Due to issues that were observed by inspection from a particular E4 sensor in the second experimental design that resulted in plausible inaccurate signal data recorded, all EDA signals obtained from that sensor in the second experimental design were removed. Specifically, that particular sensor was used by a total of 14 participants in the second experimental design; hence, the EDA signals from those 14 participants were removed from the subsequent analysis.

The remaining segmented EDA signals were pre-processed with a low-pass filter (smoothing size N = 32 samples). This helped to de-noise the signals.

4.2.3 Feature extraction

From the EDA data, the following measures were extracted:

- Mean SCR: The mean of the EDA signal. The unit was microSiemens.
- SCR Amplitude: The amplitude of the EDA signal. The unit was microSiemens.
- SCR Frequency: The frequency of the EDA signal (number of peaks of the signal per minute). The unit was peaks per minute.

From the HR data, the following measures were extracted:

- Mean HR: The mean of the HR signal. The unit was beats per minute (bpm).
- Standard Deviation of HR: The standard deviation of the HR signal. The unit was bpm.

After completion of EDA and HR signal processing, the final sample sizes (equivalent to the number of participants) in the EDA and HR datasets for the first experimental design were 24 each, while for the second experimental design they were 15 in the EDA dataset and 29 in the HR dataset. This is due to the elimination of the 14 EDA signals due to sensor issues as previously discussed; no other signals were removed during pre-processing.

The subsequent research analysis of this thesis was conducted on the final samples of participants with the aforementioned sample sizes.

4.3 Investigating impact of background discussions on stress and performance

The following results quoted in subsection 4.1 were computed for all participants, separately for each experimental design:

1. Means and standard deviations of the following for each intelligibility condition:
   (a) self-reported measures
   (b) task performance measures

    (c) EDA and HR measures

2. Task-wise correlations of the following:
    (a) self-reported measures as a function of physiological measures
    (b) task performance measures as a function of physiological measures
    (c) task performance measures as a function of self-reported measures

3. Task-wise means and standard deviations of the following for each intelligibility condition:
    (a) self-reported measures
    (b) physiological measures

We subsequently analyze above results in order to help understand the impact that background discussions have on stress and work performance, as well as to help answer the research questions of this thesis, based on insights from the patterns observed from the analyses. It is important to mention that the standard deviation results delineated above will not be included in the thesis analysis, though they will be reported in the thesis along with the other results.

It must be mentioned here that, in the above-mentioned analyses, by task-wise it is meant that the means and correlations were computed per type of cognitive task. Hence going forward in this thesis, the term task-wise will be meant in so manner.

All the correlations computed in this thesis as delineated above are Spearman's correlations.

It must also be mentioned here that, in the process of computing the results delineated above that pertain to the self-reported measures, except that of the task-wise means and standard deviations of the self-reported measures for each of the intelligibility conditions, the State Anxiety self-reported measure used was obtained as the absolute difference between the self-reported measure of State Anxiety provided before a particular cognitive task and the corresponding self-reported measure of State Anxiety provided after the cognitive task. In the process of computing the task-wise means and standard deviations of the self-reported measures for each of the intelligibility conditions, the State Anxiety self-reported measure used was the self-reported measure of State

Anxiety provided after the cognitive task.

Since we aim to determine the effect of the different speech intelligibility conditions per task type as a part of the thesis research as well, the individual Proofreading and Catalog sub-task total scores were aggregated to give a single score for the Proofreading/Catalog task (by first standardizing them to give them equal weights and then adding them together), and this aggregated score, instead of the individual sub-task scores, was used in computing the results delineated above that pertain to the task performance measures.

4.3.1 Hypotheses of the research questions

Based on the discussion until now, including based on this thesis's review of prior work, the following hypotheses were formulated for the research questions of this thesis:

1. Is distraction related to stress?
   - Increase in distraction causes increase in stress.
2. Is stress related to performance?
   - Increase in stress causes decrease in performance.
3. Is distraction related to performance?
   - Distraction and performance are inversely related (from the hypotheses of the previous two research questions).
4. Does it matter how stress is operationalized or measured between the self-reported subjective ratings and the physiological signals?
   - Based on the literature review discussed in this thesis, it is hypothesized that physiological signals are a more accurate indicator of stress than self-reported subjective ratings.
5. Does the specific type of task make a difference in any of the above results?

- The type of task is hypothesized in this thesis to have an effect on stress and performance, and thus on the above results. Correlations between each cognitive task and GMA were calculated, and it was found that the Proofreading/Catalog task overall gave the highest positive correlations with GMA, followed by the Presentation task, and then the Prioritizing task which gave the worst correlations. Based on this finding it is hypothesized in this thesis that the Proofreading/Catalog task will give results most in line with the hypotheses for the earlier results, followed by the Presentation task, and then the Prioritizing task.

6. Which intelligibility conditions have a stronger effect on stress and performance?

    - It is premised in this thesis that higher speech intelligibility corresponds to greater distraction. Thus, the high intelligibility condition will yield higher stress and lower performance compared to the low intelligibility and silence conditions.

7. Are these results consistent across different design variations of the experiment?

    - Due to the inherent definitions of the two experimental designs, the second experimental design, namely, Between-Groups (randomized), is hypothesized to be methodologically better than the first experimental design, namely, Within-Subjects, as the former is more interpretable, and therefore the Between-Groups (randomized) design variation is expected to yield results more in line with above hypotheses.

5. RESULTS

5.1 Means of the physiological signals, self-reported measures, and task performance measures

As discussed in the previous section, the means and standard deviations of the physiological signal measures, the self-reported measures, and the task performance measures across all participants in each experimental design were calculated for each intelligibility condition. All the results in this thesis are presented for the following physiological signal measures, namely, Mean SCR, SCR Amplitude, and SCR Frequency for EDA, and Mean HR and Standard Deviation of HR for HR, for the following self-reported measures, namely, State Anxiety, Trait Anxiety, Discomfort, Pop-up Stress, and cognitive workload from NASA-TLX, and for the following task performance measures, namely, Proofreading/Catalog total score, Presentation total score, and Prioritizing total absolute score. The codes of these task performance measures, as included in the table column headings in the relevant results in this thesis, are, respectively: Proofread/Catalog_Tot, Present_Tot, and Prior_Abs_Tot.

The ranges of values of each of the physiological measures, self-reported measures, and task performance measures used in computing all the results in this thesis, for each experimental design, are provided in Tables 5.1, 5.2, and 5.3, respectively. (The total number of values for which each range corresponds to is the corresponding sample size N, which will be provided in the relevant tables for each of the results in this thesis.) The values of the measures are unchanged from how they were obtained from the user study. As implied from the discussion in the previous section, two measures of the State Anxiety self-reported measure were used in computing results in this thesis, one obtained as the absolute difference between the State Anxiety measure provided before and after a cognitive task, and the other the State Anxiety measure provided after a cognitive task. The codes of these two State Anxiety measures will be assigned as State Anxiety (diff.) and State Anxiety (post), respectively, and these codes are included in the column headings of the aforementioned tables for the columns pertaining to the State Anxiety measures they respectively correspond to.

Exp. Design	Mean SCR	SCR Amplitude	SCR Frequency	Mean HR	Std. Dev. of HR
1st	0.02 - 1.15	0.007 - 0.19	0 - 4.87	63.19 - 110.47	2.02 - 24.3
2nd	0.02 - 2.13	0.007 - 0.13	0 - 3.5	68.58 - 92.99	2.12 - 20.58

Table 5.1: Ranges (approx.) of dataset values of physiological measures in each experimental design.

Exp. Design	State Anx. (diff.)	State Anx. (post)	Trait Anx.	Discomf.	Pop-up Str.	NASA-TLX
1st	-46 - 13	20 - 70	25 - 68	1 - 7	1.4 - 6.8	-95 - 495
2nd	-40 - 22	20 - 76	24 - 72	1 - 7	1.2 - 6.8	-83 - 402

Table 5.2: Ranges of dataset values of self-reported measures in each experimental design.

Exp. Design	Proofread/Catalog_Tot	Present_Tot	Abs_Basket_Tot
1st	0.21 - 1.5	3 - 14	0 - 4
2nd	0.15 - 2	4 - 17	0 - 6

Table 5.3: Ranges of dataset values of task performance measures in each experimental design.

The mean and standard deviation values of the EDA measures for each intelligibility condition for the first and second experimental design are shown in Tables 5.4 and 5.5, respectively, and the mean and standard deviation values of the HR measures for each intelligibility condition for the first and second experimental design are shown in Tables 5.6 and 5.7, respectively. The relevant sample sizes are provided in the tables as well. For the first experimental design, sample size N = 24 for each intelligibility condition for both the EDA and HR measures. For the second experimental design, N = 5 for the high intelligibility condition, N= 3 for the low intelligibility condition, and N = 7 for the silence condition for the EDA measures, while for the HR measures, N = 11 for the high intelligibility condition, N = 6 for the low intelligibility condition, and N = 12 for the silence condition. In both experimental designs for the EDA measures, the means increase overall from the high intelligibility condition to the silence condition, with the silence condition having the greatest means overall. However, in the first design variation for the HR measures, we see that the means decrease overall from the high to the silence condition, with the silence condition having the lowest means. In the second design variation, Mean HR is the highest in the silence condition, whereas Standard Deviation of HR is highest in the low intelligibility condition. Thus, only the means of the HR measures in the first design variation show patterns that are in line with our hypothesis that increase in distraction causes increase in stress.

It is important to mention here that, while analyzing the variations across intelligibility conditions in the results of this thesis, we will also be using extreme contrast as a hypothesis; thus, our extreme contrast hypothesis will be that the high intelligibility condition will yield higher stress and lower performance than will the silence condition.

In both the first and second experimental designs, considering both the EDA and HR measures together, we see that, overall, mean values are higher in the silence condition than in the high intelligibility condition across the measures. This is not in line with our extreme contrast hypothesis as well as our hypothesis that increase in distraction (intelligibility) causes increase in stress.

The mean and standard deviation values of the self-reported measures for each intelligibility condition for the first and second experimental design are shown in Tables 5.8 and 5.9, respectively, along with the relevant sample sizes. Here as well, the sample size N = 24 for each intelligibility condition for the first experimental design, while for the second experimental design, N = 11 for the high intelligibility condition, N = 6 for the low intelligibility condition, and N = 12 for the silence condition. (It is important to mention that the code, NASA-TLX, included in the column headings of the aforementioned tables refers to the cognitive workload measure from the NASA-TLX assessment. This is also the case for all the tabular results in this thesis that contain the code, NASA-TLX, either as a column heading or a row heading.) During the first experimental design, cognitive load (i.e, as measured by NASA-TLX), discomfort, and state anxiety (i.e., as measured by STAI-State) are reported the highest during the silence condition, a finding which is in contrast to our expected hypothesis (i.e., higher levels of stress, discomfort, and cognitive load are expected in noisy environments). A potential reason for this might be the fact that both the type of task and intelligibility were varying as part of the first experimental design, therefore these results might be confounded by the varying demand levels of the cognitive tasks. During the second experimental design, cognitive load and stress levels (i.e., captured by both STAI-State and the pop-up questions) are found to be the lowest during the low intelligibility condition, as compared to the silence and high intelligibility conditions. This finding is potentially in line with prior work that demonstrates that while high speech intelligibility is more detrimental to human performance [16], [17], excessive noise absorption might cause reduced acceptability to sporadic noises and also hinder performance [18]. Lastly, it is important to mention that, in general, while trying to look for patterns in stress or performance across the intelligibility conditions or across tasks, trait anxiety is relatively not very relevant as it is a trait measure and does not change over short periods of time. Thus we ignore analyzing trait anxiety in this thesis.

In the first experimental design (Table 5.8), we see that the extreme contrast hypothesis is reflected in the means of only one measure, namely pop-up stress. In the second experimental design (Table 5.9), however, the extreme contrast hypothesis is reflected in the means of half of the measures, namely, pop-up stress and cognitive load.

Similarly, the means and standard deviations of the task performance measures for each intelligibility condition for the first and second experimental design are reported in Tables 5.10 and 5.11, along with the relevant sample sizes. Here, the sample size N = 8 for each intelligibility condition for the first experimental design, while for the second experimental design, N = 11 for the high intelligibility condition, N = 6 for the low intelligibility condition, and N = 12 for the silence condition. During the first experimental design, Prioritizing total absolute score is reported to be the highest during the silence condition and the lowest during the low intelligibility condition. Proofreading/Catalog total score and Presentation total score are both reported to be the highest during the high intelligibility condition, while Presentation total score is reported to be the lowest during the silence condition. Thus, in the first design variation, only the Prioritizing total absolute score shows a pattern in line with our extreme contrast hypothesis. However, we see an overall increase in the means of the Prioritizing total absolute score from the high intelligibility condition to the silence condition, which is in line with our hypothesis that increase in distraction cause decrease in performance. During the second experimental design, only Presentation total score is reported to be the highest during the silence condition. Performance scores were found the lowest in the silence condition for the Proofreading/Catalog task. Thus, only the patterns we see in the Presentation total score are in line with our extreme contrast hypothesis as well as our hypothesis that increase in distraction cause decrease in performance. An overall decrease in the means of the Prioritizing total absolute score shows that this pattern is in line with the latter hypothesis. The accuracy of the results from a Between-Groups (randomized) design is dependent on the efficacy of the random assignment used, a factor not present in a Within-Subjects design. This is a possible explanation for the findings from the second experimental design that are not in line with our hypothesis that increase in distraction causes decrease in performance.

Intelligibility Cond.	**Mean SCR**	**SCR Amplitude**	**SCR Frequency**
high [24]	0.14, 0.12	0.02, 0.01	0.77, 1.01
low [24]	0.15, 0.13	0.02, 0.01	0.92, 1.23
none [24]	0.2, 0.25	0.03, 0.05	0.87, 1.25

Table 5.4: Means and standard deviations of EDA measures across the three speech intelligibility conditions in the first experimental design.

Intelligibility Cond.	Mean SCR	SCR Amplitude	SCR Frequency
high [5]	0.15, 0.06	0.03, 0.02	0.89, 1
low [3]	0.27, 0.25	0.04, 0.04	1.53, 1.31
none [7]	0.45, 0.67	0.04, 0.03	1.64, 1.14

Table 5.5: Means and standard deviations of EDA measures across the three speech intelligibility conditions in the second experimental design.

Intelligibility Cond.	Mean HR	Standard Deviation of HR
high [24]	77.25, 6.43	9.08, 4.19
low [24]	78.43, 10.51	9.04, 6
none [24]	77.1, 7.75	8.04, 4.67

Table 5.6: Means and standard deviations of HR measures across the three speech intelligibility conditions in the first experimental design.

Intelligibility Cond.	Mean HR	Standard Deviation of HR
high [11]	78.59, 6.06	6.87, 4.8
low [6]	78.52, 7.17	9.63, 4.86
none [12]	79.64, 4.82	8.57, 4.34

Table 5.7: Means and standard deviations of HR measures across the three speech intelligibility conditions in the second experimental design.

Intelligibility Cond.	State Anxiety	Trait Anxiety	Discomf.	Pop-up Stress	NASA-TLX
high [24]	6, 11.12	45.67, 9.71	3.04, 1.57	4.37, 1.33	145.58, 140.91
low [24]	5.67, 12.71	45.67, 9.71	3.38, 1.72	4.24, 1.44	171.5, 155
none [24]	10.79, 10.33	45.67, 9.71	3.67, 1.44	4.13, 1.18	173.29, 116.74

Table 5.8: Means and standard deviations of self-reported ratings across the three speech intelligibility conditions in the first experimental design.

Intelligibility Cond.	State Anxiety	Trait Anxiety	Discomf.	Pop-up Stress	NASA-TLX
high [11]	6.03, 8.43	41.27, 9.39	3.52, 1.5	4.04, 1.5	175.3, 122.06
low [6]	4.28, 9.23	46, 14.95	3.95, 1.47	3.49, 1.16	117.22, 124.62
none [12]	6.19, 12.15	44.58, 10.81	3.53, 1.5	4.03, 1.11	132.17, 108.12

Table 5.9: Means and standard deviations of self-reported ratings across the three speech intelligibility conditions in the second experimental design.

Intelligibility Cond.	Proofread/Catalog_Tot	Present_Tot	Abs_Basket_Tot
high [8]	1.01, 0.34	10.19, 1.98	1.75, 1.04
low [8]	0.63, 0.37	9.38, 4.21	1.57, 1.27
none [8]	0.91, 0.37	9.19, 2.12	2.88, 1.13

Table 5.10: Means and standard deviations of task performance measures across the three speech intelligibility conditions in the first experimental design.

Intelligibility Cond.	Proofread/Catalog_Tot	Present_Tot	Abs_Basket_Tot
high [11]	1.04, 0.5	9.41, 3.27	1.64, 1.86
low [6]	1.07, 0.5	9.5, 4.1	3.33, 1.75
none [12]	0.82, 0.47	11.79, 2.96	1.83, 1.4

Table 5.11: Means and standard deviations of task performance measures across the three speech intelligibility conditions in the second experimental design.

5.2 Task-wise correlations between the physiological measures and the self-reported measures

The task-wise correlations between the physiological measures of all participants and their self-reported measures were computed for each experimental design as well. Sample size N = 24 per task for the first experimental design with the EDA measures and the HR measures each, while for the second experimental design, N = 15 per task with the EDA measures and N = 29 per task with the HR measures. Correlations were computed between each self-reported measure for each task and the corresponding EDA and HR physiological measure for that task.

The task-wise correlation values between the EDA measures and the self-reported measures for the first and second experimental design are shown in Tables 5.12 and 5.13, respectively. The task-wise correlation values between the HR measures and the self-reported measures for the first and second experimental design are shown in Tables 5.14 and 5.15, respectively. The corresponding sample sizes as aforementioned are reported in the respective tables as well. In the tables, the '1, 2, 3' next to the name of each self-reported measure in the column headings represents the

Proofreading/Catalog task, the Presentation task, and the Prioritizing task, respectively; thus each column heading denotes one of the self-reported measures for each of the three task types.

In the first experimental design, considering both the EDA and HR measures together, we see an overall positive correlation with all self-reported measures. We also see a majority of relatively high-magnitude negative correlation values among the negative correlation values in the first design variation. In the second experimental design, considering both the EDA and HR measures together, we see an overall positive correlation with all self-reported measures (we see an overall negative correlation with discomfort). We also see here that most of the negative correlation values with all self-reported measures are relatively low-magnitude. Thus, the correlation values in the second experimental design suggest more than those in the first experimental design that both physiological measures and self-reported measures are equivalently indicative of stress.

When analysing the results task-wise, we see that, in the first experimental design, the cases where we do not see a preponderance of positive and relatively low magnitude negative correlation values are the correlations of the physiological measures with pop-up stress in the Proofreading/Catalog and Presentation tasks. In the second experimental design, we do not see any individual cases where we do not see a preponderance of positive and relatively low magnitude negative correlation values.

EDA Meas.	**State Anxiety 1, 2, 3**	**Trait Anxiety 1, 2, 3**	**Discomf. 1, 2, 3**	**Pop-up Stress 1, 2, 3**	**NASA-TLX 1, 2, 3**
Mean SCR [24 per task]	0.41, 0.06, -0.08	-0.04, -0.04, -0.09	0.43, 0.29, 0.17	-0.31, -0.12, -0.04	0.43, 0.38, -0.07
SCR Amplitude [24 per task]	0.31, -0.2, -0.02	-0.45, -0.22, -0.24	0.47, 0.17, 0.04	-0.32, -0.46, 0.31	0.37, -0.13, -0.09
SCR Frequency [24 per task]	0.35, 0.32, -0.17	-0.13, -0.14, -0.02	0.43, 0.37, 0.29	-0.36, -0.28, -0.14	0.4, 0.46, 0.04

Table 5.12: Task-wise correlations between EDA measures of the participants and their self-reported measures in the first experimental design.

EDA Meas.	State Anxiety 1, 2, 3	Trait Anxiety 1, 2, 3	Discomf. 1, 2, 3	Pop-up Stress 1, 2, 3	NASA-TLX 1, 2, 3
Mean SCR [15 per task]	0.23, -0.14, -0.02	0.04, -0.09, -0.09	-0.18, -0.04, -0.09	0.16, 0.57, 0.21	0.19, 0.11, -0.03
SCR Amplitude [15 per task]	0.02, 0.05, -0.15	0.07, -0.31, 0.03	-0.27, -0.04, 0.01	0.26, 0.34, 0.19	0.03, 0.32, 0.01
SCR Frequency [15 per task]	0.23, 0.07, 0.1	0.03, -0.21, -0.15	-0.13, -0.07, -0.17	-0.05, 0.3, 0.32	0.15, 0.29, -0.06

Table 5.13: Task-wise correlations between EDA measures of the participants and their self-reported measures in the second experimental design.

HR Meas.	State Anxiety 1, 2, 3	Trait Anxiety 1, 2, 3	Discomf. 1, 2, 3	Pop-up Stress 1, 2, 3	NASA-TLX 1, 2, 3
Mean HR [24 per task]	-0.05, -0.13, -0.14	-0.2, -0.01, -0.27	-0.25, -0.33, -0.37	0.09, 0.08, 0.52	-0.11, 0.16, -0.27
Std. Dev of HR [24 per task]	0.15, -0.13, -0.27	-0.08, 0.1, -0.12	0.14, -0.23, -0.29	-0.23, -0.27, 0.31	0.15, 0.36, -0.03

Table 5.14: Task-wise correlations between HR measures of the participants and their self-reported measures in the first experimental design.

HR Meas.	State Anxiety 1, 2, 3	Trait Anxiety 1, 2, 3	Discomf. 1, 2, 3	Pop-up Stress 1, 2, 3	NASA-TLX 1, 2, 3
Mean HR [29 per task]	0.53, 0.22, -0.08	-0.19, 0.02, -0.24	0.08, 0.06, 0.11	-0.05, -0.34, 0.02	0.24, 0.09, -0.1
Std. Dev of HR [29 per task]	0.18, 0.29, -0.18	-0.12, 0.03, -0.1	0.24, -0.11, -0.19	-0.22, -0.03, -0.01	0.16, 0.1, -0.25

Table 5.15: Task-wise correlations between HR measures of the participants and their self-reported measures in the second experimental design.

5.3 Task-wise correlations between the physiological measures and the task performance measures

The task-wise correlations between the physiological measures of all participants and their task performance measures were computed for each experimental design as well. Sample size N = 24 for the first experimental design with the EDA measures and the HR measures each, while for the second experimental design, N = 15 with the EDA measures and N = 29 with the HR measures. Correlations were computed between each task performance measure and the corresponding EDA and HR physiological measure for that task, namely, Proofreading/Catalog, Presentation, and Prioritizing.

The task-wise correlation values between the EDA measures and the task performance measures for the first and second experimental design are shown in Tables 5.16 and 5.17, respectively. The task-wise correlation values between the HR measures and the task performance measures for the first and second experimental design are shown in Tables 5.18 and 5.19, respectively. The corresponding sample sizes as aforementioned are reported in the respective tables as well. Based on our hypothesis that increase in stress causes decrease in performance, we would expect more negative correlation values and relatively low magnitude positive correlation values. Considering both tables for the EDA and HR measures for the first experimental design together, we see that there are more negative and relatively low magnitude positive correlation values for all tasks overall. Considering both tables for the EDA and HR measures for the second experimental design together, we see that, again, there are more negative and relatively low magnitude positive correlation values for all tasks overall. Thus these findings from both design variations are in line with our hypothesis.

EDA Meas.	**Proofread/Catalog_Tot**	**Present_Tot**	**Prior_Abs_Tot**
Mean SCR [24 per task]	-0.17	0.02	-0.09
SCR Amplitude [24 per task]	-0.22	0.03	-0.36
SCR Frequency [24 per task]	-0.25	0.13	-0.15

Table 5.16: Task-wise correlations between EDA measures of the participants and their task performance measures in the first experimental design.

EDA Meas.	**Proofread/Catalog_Tot**	**Present_Tot**	**Prior_Abs_Tot**
Mean SCR [15 per task]	-0.25	0.09	0.0
SCR Amplitude [15 per task]	-0.21	0.27	-0.49
SCR Frequency [15 per task]	-0.22	0.06	-0.48

Table 5.17: Task-wise correlations between EDA measures of the participants and their task performance measures in the second experimental design.

HR Meas.	Proofread/Catalog_Tot	Present_Tot	Prior_Abs_Tot
Mean HR [24 per task]	-0.05	-0.31	0.5
Std. Dev. of HR [24 per task]	-0.19	-0.28	0.42

Table 5.18: Task-wise correlations between HR measures of the participants and their task performance measures in the first experimental design.

HR Meas.	Proofread/Catalog_Tot	Present_Tot	Prior_Abs_Tot
Mean HR [29 per task]	0.13	-0.03	-0.28
Std. Dev. of HR [29 per task]	0.12	0.15	0.18

Table 5.19: Task-wise correlations between HR measures of the participants and their task performance measures in the second experimental design.

5.4 Task-wise correlations between the self-reported measures and the task performance measures

The task-wise correlations between the self-reported measures of all participants and their task performance measures were computed for each experimental design as well. Sample size N = 24 for the first experimental design, while N = 29 for the second experimental design. Correlations were computed between each self-reported measure for each task, namely, Proofreading/Catalog, Presentation, and Prioritizing, and the corresponding task performance measure.

The task-wise correlation values for the first and second experimental design are shown in Tables 5.20 and 5.21, respectively, along with the corresponding aforementioned sample sizes. Based on our hypothesis that increase in stress causes decrease in performance, we would expect more negative correlation values and relatively low magnitude positive correlation values. In the first design variation, we see a preponderance of negative correlation values and relatively low magnitude positive correlation values, which is in line with our hypothesis. In the second design variation, we see a similar scenario, thus in line with our hypothesis; however, we see more negative correlation values than in the first design variation.

Self-rep. Sub. Rating	Proofread/Catalog_Tot	Present_Tot	Abs_Basket_Tot
State Anx. [24 per task]	0.12	0.14	0.29
Trait Anx. [24 per task]	-0.08	-0.25	-0.2
Discomf. [24 per task]	-0.12	0.46	0.01
Pop-up Stress [24 per task]	0.26	-0.32	-0.07
NASA-TLX [24 per task]	-0.09	0.02	0.14

Table 5.20: Task-wise correlations between self-reported measures of the participants and their task performance measures in the first experimental design.

Self-rep. Sub. Rating	Proofread/Catalog_Tot	Present_Tot	Abs_Basket_Tot
State Anxiety [29 per task]	-0.22	0.1	-0.07
Trait Anxiety [29 per task]	-0.16	-0.09	0.13
Discomfort [29 per task]	-0.08	0.06	-0.02
Pop-up Stress [29 per task]	-0.06	0.28	-0.19
NASA-TLX [29 per task]	-0.26	-0.4	-0.13

Table 5.21: Task-wise correlations between self-reported measures of the participants and their task performance measures in the second experimental design.

5.5 Task-wise means and standard deviations of the physiological measures per intelligibility condition

Means and standard deviations of the physiological measures across all participants were computed for each intelligibility condition per task as well. Sample size N = 8 per intelligibility condition per task for the first experimental design for the EDA measures and the HR measures each, while for the second experimental design, N = 5 per task for the high intelligibility condition, N = 3 per task for the low intelligibility condition, and N = 7 per task for the silence condition for the EDA measures, and N = 11 per task for the high intelligibility condition, N = 6 per task for the low intelligibility condition, and N = 12 per task for the silence condition for the HR measures. Means and standard deviations of each EDA and HR physiological measure were computed for each intelligibility condition per task.

The task-wise mean values for the EDA measures per intelligibility condition for the first and second experimental design are shown in Tables 5.22 and 5.24, respectively, while the task-wise standard deviation values for the EDA measures per intelligibility condition for the first and second experimental design are shown in Tables 5.23 and 5.25, respectively. The task-wise mean values per intelligibility condition for the HR measures for the first and second experimental design are shown in Tables 5.26 and 5.28, respectively, while the task-wise standard deviation values per intelligibility condition for the HR measures for the first and second experimental design are shown in Tables 5.27 and 5.29, respectively. The corresponding sample sizes as aforementioned are reported in the respective tables as well. In the tables, the '1, 2, 3' next to the name of each EDA and HR physiological measure in the column headings represents the Proofreading/Catalog task, the Presentation task, and the Prioritizing task, respectively; thus each column heading denotes one of the physiological measures for each of the three task types.

In the first experimental design, considering both the EDA and HR measures together, we see that the mean values increase overall from the high intelligibility condition to the silence condition. In the second experimental design, again considering both the EDA and HR measures together, the mean values increase overall from the high intelligibility condition to the silence condition across the measures here as well. These findings from both design variations are not in line with our hypothesis that increase in distraction causes increase in stress, nor are they in line with our extreme contrast hypothesis. However, considering the EDA and HR measures separately, we see that, in the first experimental design, the mean values of only the EDA measures and Standard Deviation of HR from the Prioritizing task, and the mean values of the HR measures from the Proofreading/Catalog task, show a pattern which is in line with our extreme contrast hypothesis, and thus our hypothesis that increase in distraction causes increase in stress. In the second experimental design, we see that the mean values of only Mean HR from the Proofreading/Catalog task show a pattern which is in line with our extreme contrast hypothesis, and thus our hypothesis that increase in distraction causes increase in stress.

Intelligibility Cond.	Mean SCR 1, 2, 3	SCR Amplitude 1, 2, 3	SCR Frequency 1, 2, 3
high [8 per task]	0.14, 0.13, 0.15	0.03, 0.02, 0.02	0.73, 0.67, 0.93
low [8 per task]	0.18, 0.13, 0.14	0.01, 0.02, 0.03	0.95, 0.69, 1.1
none [8 per task]	0.25, 0.2, 0.13	0.06, 0.02, 0.01	1.04, 0.82, 0.71

Table 5.22: Task-wise means of EDA measures of the participants for each intelligibility condition in the first experimental design.

Intelligibility Cond.	Mean SCR 1, 2, 3	SCR Amplitude 1, 2, 3	SCR Frequency 1, 2, 3
high [8 per task]	0.14, 0.13, 0.1	0.02, 0.01, 0.01	1.16, 0.93, 1.08
low [8 per task]	0.15, 0.11, 0.15	0.01, 0.01, 0.01	0.98, 1.22, 1.54
none [8 per task]	0.39, 0.17, 0.12	0.09, 0.01, 0.01	1.66, 1.01, 1.08

Table 5.23: Task-wise standard deviations of EDA measures of the participants for each intelligibility condition in the first experimental design.

Intelligibility Cond.	Mean SCR 1, 2, 3	SCR Amplitude 1, 2, 3	SCR Frequency 1, 2, 3
high [5 per task]	0.17, 0.16, 0.11	0.03, 0.03, 0.02	1.05, 0.5, 1.11
low [3 per task]	0.19, 0.17, 0.43	0.04, 0.03, 0.05	1.52, 1.15, 1.92
none [7 per task]	0.36, 0.53, 0.46	0.04, 0.03, 0.04	1.58, 1.42, 1.92

Table 5.24: Task-wise means of EDA measures of the participants for each intelligibility condition in the second experimental design.

Intelligibility Cond.	Mean SCR 1, 2, 3	SCR Amplitude 1, 2, 3	SCR Frequency 1, 2, 3
high [5 per task]	0.04, 0.05, 0.07	0.02, 0.02, 0.01	0.85, 0.42, 1.57
low [3 per task]	0.06, 0.12, 0.42	0.02, 0.02, 0.06	1.22, 1.86, 1.2
none [7 per task]	0.44, 0.81, 0.78	0.03, 0.02, 0.04	1.22, 1.33, 1.01

Table 5.25: Task-wise standard deviations of EDA measures of the participants for each intelligibility condition in the second experimental design.

Intelligibility Cond.	Mean HR 1, 2, 3	Standard Deviation of HR 1, 2, 3
high [8 per task]	80.48, 73.49, 77.77	9.89, 7.04, 10.3
low [8 per task]	81.66, 79.94, 73.69	7.4, 10.69, 9.03
none [8 per task]	74.36, 76.18, 80.75	5.96, 8.5, 9.67

Table 5.26: Task-wise means of HR measures of the participants for each intelligibility condition in the first experimental design.

Intelligibility Cond.	**Mean HR 1, 2, 3**	**Standard Deviation of HR 1, 2, 3**
high [8 per task]	6.46, 4.21, 6.94	5.42, 2.28, 3.99
low [8 per task]	14.13, 9.24, 6.16	6.14, 5.26, 6.83
none [8 per task]	4.51, 7.86, 9.5	3.27, 4.54, 5.65

Table 5.27: Task-wise standard deviations of HR measures of the participants for each intelligibility condition in the first experimental design.

Intelligibility Cond.	**Mean HR 1, 2, 3**	**Standard Deviation of HR 1, 2, 3**
high [11 per task]	80.35, 78.81, 76.76	6.25, 7.78, 6.52
low [6 per task]	77.99, 79.59, 77.99	8.42, 10.29, 10.19
none [12 per task]	79.33, 79.78, 79.79	7.84, 9.64, 8.16

Table 5.28: Task-wise means of HR measures of the participants for each intelligibility condition in the second experimental design.

Intelligibility Cond.	**Mean HR 1, 2, 3**	**Standard Deviation of HR 1, 2, 3**
high [11 per task]	7.39, 6.68, 3.77	4.81, 5.65, 4.22
low [6 per task]	5.81, 8.49, 8.19	3.62, 6.41, 4.86
none [12 per task]	5.25, 5.41, 4.23	3.85, 4.81, 4.48

Table 5.29: Task-wise standard deviations of HR measures of the participants for each intelligibility condition in the second experimental design.

5.6 Task-wise means and standard deviations of the self-reported measures per intelligibility condition

Means and standard deviations of the self-reported measures across all participants were computed for each intelligibility condition per task as well. Sample size N = 8 per intelligibility condition task for the first experimental design, while for the second experimental design, N = 11 per task for the high intelligibility condition, N = 6 per task for the low intelligibility condition, and N = 12 per task for the silence condition. Means and standard deviations of each self-reported measure were computed for each intelligibility condition per task.

The task-wise mean values for the self-reported measures per intelligibility condition for the first and second experimental design are shown in Tables 5.30 and 5.32, respectively. The task-wise standard deviation values for the self-reported measures for the first and second experimental design are shown in Tables 5.31 and 5.33, respectively. The corresponding sample sizes as aforementioned are reported in the respective tables as well. In the tables, the '1, 2, 3' next to the name of each self-reported measure in the column headings represents the Proofreading/Catalog task, the Presentation task, and the Prioritizing task, respectively; thus each column heading denotes one of the self-reported measures for each of the three task types.

In the first experimental design, we see that overall, the means of the self-reported measures increase from the high intelligibility condition to the silence condition. We see a similar scenario in the second experimental design. These findings from both design variations are not in line with our hypothesis that increase in distraction causes increase in stress, nor are they in line with our extreme contrast hypothesis. When analysing the results task-wise, we see that, in the first experimental design, the mean values of only state anxiety, discomfort, and cognitive load (from NASA-TLX) from the Presentation task, and the mean values of only pop-up stress from the Proofreading/Catalog and Prioritizing tasks, show a pattern which is in line with our extreme contrast hypothesis, and thus our hypothesis that increase in distraction causes increase in stress. In the second experimental design, the mean values of only discomfort, pop-up stress, and cognitive load from the Prioritizing task, the mean values of pop-up stress and discomfort from the Presentation task, and the mean values of cognitive load from the Proofreading/Catalog task, show a pattern in line with our extreme contrast hypothesis, and thus our hypothesis that increase in distraction cause increase in stress.

Intelligibility Cond.	State Anxiety 1, 2, 3	Trait Anxiety 1, 2, 3	Discomf. 1, 2, 3	Pop-up Stress 1, 2, 3	NASA-TLX 1, 2, 3
high [8 per task]	36.88, 44.5, 27.88	48.25, 45.5, 43.25	2.75, 4.38, 2.0	4.6, 3.23, 5.28	126.5, 273.88, 36.38
low [8 per task]	35.38, 47.88, 34.25	43.25, 48.25, 45.5	2.75, 4.38, 3.0	4.37, 3.7, 4.65	111.5, 337.88, 65.13
none [8 per task]	41.0, 41.0, 40.5	45.5, 43.25, 48.25	3.5, 4.13, 3.38	4.35, 3.57, 4.4	166.88, 237.13, 115.88

Table 5.30: Task-wise means of self-reported measures of the participants for each intelligibility condition in the first experimental design.

Intelligibility Cond.	State Anxiety 1, 2, 3	Trait Anxiety 1, 2, 3	Discomf. 1, 2, 3	Pop-up Stress 1, 2, 3	NASA-TLX 1, 2, 3
high [8 per task]	11.91, 13.23, 4.55	9.25, 13.25, 5.87	1.04, 1.85, 0.54	1.08, 1.17, 0.86	128.36, 82.72, 95.2
low [8 per task]	5.95, 11.26, 12.52	5.87, 9.25, 13.25	1.17, 1.92, 1.69	1.45, 1.58, 1.31	77.44, 88.66, 128.05
none [8 per task]	11.8, 8.32, 14.59	13.25, 5.87, 9.25	1.69, 1.25, 1.41	1.23, 1.22, 1.08	134.55, 102.52, 87.73

Table 5.31: Task-wise standard deviations of self-reported measures of the participants for each intelligibility condition in the first experimental design.

Intelligibility Cond.	State Anxiety 1, 2, 3	Trait Anxiety 1, 2, 3	Discomf. 1, 2, 3	Pop-up Stress 1, 2, 3	NASA-TLX 1, 2, 3
high [11 per task]	32.1, 39.36, 37.73	41.27, 41.27, 41.27	2.8, 4.18, 3.46	4.2, 3.54, 4.6	113.0, 248.82, 160.18
low [6 per task]	39.67, 47.83, 41.33	46.0, 46.0, 46.0	3.33, 4.67, 3.83	4.08, 2.48, 3.92	77.33, 226.83, 47.5
none [12 per task]	35.83, 43.58, 39.33	44.58, 44.58, 44.58	3.0, 4.25, 3.33	4.33, 3.38, 4.42	79.5, 216.5, 100.5

Table 5.32: Task-wise means of self-reported measures of the participants for each intelligibility condition in the second experimental design.

Intelligibility Cond.	**State Anxiety 1, 2, 3**	**Trait Anxiety 1, 2, 3**	**Discomf. 1, 2, 3**	**Pop-up Stress 1, 2, 3**	**NASA-TLX 1, 2, 3**
high [11 per task]	11.12, 10.95, 7.98	9.7, 9.7, 9.7	1.57, 1.54, 1.14	1.39, 1.72, 1.23	124.6, 119.93, 90.41
low [6 per task]	10.21, 15.68, 16.99	15.91, 15.91, 15.91	1.63, 0.82, 1.72	0.68, 0.92, 1.08	131.73, 70, 94.39
none [12 per task]	11.8, 12.96, 13.96	11.13, 11.13, 11.13	1.12, 1.49, 1.62	1.1, 1.02, 0.94	82.12, 72.49, 116.35

Table 5.33: Task-wise standard deviations of self-reported measures of the participants for each intelligibility condition in the second experimental design.

6. DISCUSSION

From the work of this thesis and the ensuing results and analyses as discussed in the previous section, we answer the aimed research questions as follows (the tables referenced subsequently are denoted in square brackets where appropriate). It is important to mention here that we will be using only our extreme contrast hypothesis while analyzing the variations across intelligibility conditions in the results of this thesis, and our findings from this will be part of the base for deriving conclusions from the results (as will be done subsequently in this section).

1. **Is distraction related to stress?**

    - Considering the hypothesized stress indicator to be only physiological measures [Tables 5.4 - 5.7]: In the second experimental design, the means of both the EDA and HR measures are inversely related overall to the intelligibility level. However, in the first experimental design, the means of the HR measures are directly related overall to the intelligibility level, while the means of the EDA measures are inversely related overall to the intelligibility level. Thus, considering the physiological measures as a whole, and considering both experimental designs, the means of the physiological measures are inversely related overall to the intelligibility level (since a greater number of physiological measures are inversely related overall to the intelligibility level).
    - Considering the hypothesized stress indicator to be only self-reported subjective stress ratings [Tables 5.8 - 5.9]: During the first experimental design, the means of state anxiety (from STAI-State), discomfort, and cognitive load (from NASA-TLX) are inversely related overall to the intelligibility level, whereas pop-up stress is directly related overall to the intelligibility level. During the second experimental design, the means of state anxiety and discomfort are inversely related overall to the intelligibility level, while the means of pop-up stress and cognitive load are directly related overall to the intelligibility level, a finding which partially conforms to our hypothesis for this research question

(since exactly half the self-reported measures are directly related overall to the intelligibility level). Since there is no majority of self-reported subjective measures whose means show an overall direct relation with the intelligibility level in either experimental design, it is concluded that, considering both experimental designs, the means of the self-reported subjective stress ratings as a whole are inversely related overall to the intelligibility level.

Thus from the above results and observations, in general, the means per intelligibility condition of both hypothesized stress indicators do not reflect our hypothesis completely. This could be due to the fact the sample sizes for which the means were calculated in both experimental designs are relatively small in general. However, we can see that in each experimental design, the means of the self-reported measures fare better in terms of conforming to our hypothesis than do the means of the physiological measures.

Thus, increase in distraction did not lead to increase in stress overall for either experimental design; however the results in the second design were closer to being in line with this hypothesis when considering self-reported subjective stress ratings as the only indicator of stress, while the results in the first design were closer to being in line with the aforementioned hypothesis when considering physiological measures as the only indicator of stress.

2. **Is stress related to performance?**

   - Considering the hypothesized stress indicator to be only physiological measures [Tables 5.16 - 5.19]: In both experimental designs, a majority of negative and relatively low magnitude positive correlations was shown overall between the task performance measures and the EDA and HR measures.

   - Considering the hypothesized stress indicator to be only self-reported subjective stress ratings [Tables 5.20 - 5.21]: In both experimental designs, a majority of negative and relatively low magnitude positive correlations was shown overall between the task performance measures and the self-reported measures.

Thus, in general, in both experimental designs, both hypothesized stress indicators show overall an inverse relation to performance, which is in line with our hypothesis for this research question.

3. **Is distraction related to performance?** [Tables 5.10 - 5.11] In the first experimental design, performance scores were overall directly related to the intelligibility level for the Proofreading/Catalog and Presentation tasks, but were inversely related in the case of the Prioritizing task. In the second experimental design, performance scores were overall directly related to intelligibility for the Proofreading/Catalog task, but were overall inversely related in the case of the Presentation and Prioritizing tasks. Thus, the Prioritizing task is the only task in both design variations that conforms to our hypothesis for this question.

   Moreover, a majority of the means of the task performance measures conforming with our hypothesis is shown only in the second experimental design, but not in the first experimental design. Thus, it is concluded that our hypothesis does not reflect in the first experimental design, but is reflected overall in the second experimental design. The relatively lower sample sizes overall for which the means of the physiological measures were calculated in the first experimental design, compared to the sample sizes overall for which the means of the self-reported measures were calculated in the same design, could be a factor; however, further analysis might need to be conducted in order to try and confirm this.

4. **Does it matter how stress is operationalized or measured between self-reported subjective ratings and physiological measures?** [Tables 5.12 - 5.15] In both experimental designs, a majority of positive and low-magnitude negative correlations was shown overall between the physiological measures and the self-reported measures. Thus the physiological measures were overall positively correlated with the self-reported measures for both design variation, based on this finding, suggesting that they are both equivalent indicators of stress response.

   However, to answer this research question, we consider additional results as well (which, if analyzing them to answer the question proves conclusive, will provide a more decisive answer to the question): the mean values per intelligibility condition of the physiological measures and of the self-reported measures [Tables 5.4 - 5.9], the task-wise correlations between the physiological measures and the task performance measures [Tables 5.16 - 5.19], and the task-wise correlations between the self-reported measures and the task performance measures [5.20 - 5.21]. Based on these results and the analyses and inferences based on them as discussed thus far, in the first experimental design the means per intelligibility condition of the physiological measures were more in line with the corresponding hypothesis than the self-reported measures, while the scenario was vice versa in the the second experimental design. The task-wise correlations of the task performance measures between both the physiological measures and the self-reported measures were overall in line with the corresponding hypothesis in both experimental designs. Thus, considering all these results and observations (as were mentioned as relevant to answering this research question), it is concluded that physiological measures are a more accurate indicator of stress than self-reported subjective stress ratings in the first experimental design, while in the second experimental design self-reported subjective ratings are a more accurate indicator of stress. The latter observation goes against our hypothesis that physiological measures are a more accurate stress indicator than self-reported subjective stress ratings, while the former observation is in line with the hypothesis.

Though the sample sizes for which the means of the HR measures were calculated per intelligibility condition in the second design are the same as the sample sizes for which the means of the self-reported measures were calculated in the same design, the sample sizes for which the means of the EDA measures were calculated in this design are relatively lower, with the EDA measures being a greater proportion of the total set of physiological measures than the HR measures, and this could be a factor in why we do not see the mean values per intelligibility condition in the second experimental design conforming to our hypothesis that physiological measures are a more accurate indicator of stress.

5. **Does the specific type of task make a difference in any of the above results?**

    - Analyzing the means per intelligibility condition of the task performance measures [Tables 5.10 - 5.11], we had seen a difference among the task types. Specifically, the means of only Prioritizing total absolute score conform to our corresponding hypothesis in the first experimental design, while in the second experimental design the means of only Presentation total score and Prioritizing total absolute score conform to the hypothesis.
    - Analyzing the task-wise correlations between the physiological measures and the self-reported measures [Tables 5.12 - 5.15], we see in both experimental designs some differences among the task types at the individual correlation value level as well as at the level of the self-reported measure (an example of the latter of which was discussed in subsection 5.2); however, further analysis on these differences might need to be conducted in order to come to a conclusion as to whether these differences are significant and/or worth considering.
    - Analyzing the task-wise correlations between the physiological measures and the task performance measures [Tables 5.16 - 5.19], we see that, in the first experimental design, the correlation values with Proofreading/Catalog total score are all negative, while the correlation values with Presentation total score are all either negative or of relatively low positive magnitude, and the correlation values with Prioritizing total absolute score only show a majority of negative and relatively low magnitude positive correlation values. In the second experimental design, the correlation values with the total scores of all the tasks are predominantly negative and relatively low magnitude positive. Thus, based on these findings, in the first experimental design the task-wise correlation values between the physiological measures and the Proofreading/Catalog total scores show a pattern most in line with the corresponding hypothesis, followed by the task-wise correlation values with the Presentation total scores, and then followed last by the task-wise correlation values with the Prioritizing total absolute scores. In the second experimen-

tal design, the task-wise correlation values are overall in line with the corresponding hypothesis for all the task types.

- Analyzing the task-wise correlations between the self-reported measures and the task performance measures [Tables 5.20 - 5.21], we see that, in the first experimental design, the correlation values with all three tasks show a majority of negative and relatively low magnitude positive correlation values, while in the second experimental design, the correlation values with Proofreading/Catalog total score are all negative, the correlation values with Prioritizing total absolute score are all either negative or of relatively low positive magnitude, and the correlation values with Presentation total score only show a majority of negative and relatively low magnitude positive correlation values. Thus, based on these findings, in the first experimental design, the task-wise correlation values are overall in line with the corresponding hypothesis for all the task types, while in the second experimental design the task-wise correlation values between the physiological measures and the Proofreading/Catalog total scores show a pattern most in line with the corresponding hypothesis, followed by the task-wise correlation values with the Prioritizing total absolute scores, and then followed last by the task-wise correlation values with the Presentation total scores.

When we look at the results and observations (as discussed as part of this research question thus far) pertaining to the means of the task performance measures per intelligibility condition, we do see differences among the three tasks, a finding which is in line with our hypothesis that the specific type of task does make a difference in the earlier results. However, we do not see our hypothesis that these results will be most in line with the corresponding hypothesis for the Proofreading/Catalog task, followed by the Presentation task, and followed last by the Prioritizing task, reflected in these results.

When we look at the results and observations (as discussed as part of this research question thus far) pertaining to the task-wise correlations between the physiological measures and the self-reported measures, we again do see differences among the three tasks (thus being in line with our hypothesis that the specific type of task does make a difference in the earlier results), but at the level of analysis done as part of this thesis.

When we look at the results and observations (as discussed as part of this research question thus far) pertaining to the task-wise correlations between the task performance measures and the physiological measures, and the task-wise correlations between the task performance measures and the self-reported measures, we see that, either it is shown obviously that the results with respect to the Proofreading/Catalog task are most in line with the corresponding hypothesis when compared to the respective results of the Presentation and Prioritizing tasks, or the results are overall in line with the corresponding hypotheses for all three tasks. Thus, it is concluded that, in the former case, the specific type of task did have an obvious effect on the corresponding results compared to the latter case, where the specific type of task did not have a comparatively obvious effect on the corresponding results. (Further analysis can be done on the individual correlation values in results pertaining to the latter case to identify differences in patterns among the three tasks.) Thus, our hypotheses for this research question are reflected in the former case, while they are not reflected in the latter case as per the level of analysis done in this thesis.

However, in order to answer this research question, we also consider the task-wise means of the physiological measures and the self-reported measures per intelligibility condition.

- From the task-wise means of the physiological measures per intelligibility condition [Tables 5.22 - 5.29], we saw that, in the first experimental design, the mean values of some of the physiological measures in the Proofreading/Catalog task and the Prioritizing task show a pattern which is in line with the hypothesis that increase in distraction causes increase in stress. In the second experimental design, we saw that only the mean values of one of the physiological measures in the Proofreading/Catalog task show a pattern which is in line with the aforementioned hypothesis. Thus, from these findings it can be seen that the Proofreading/Catalog and Prioritizing tasks show results more in line with the aforementioned hypothesis in the first experimental design, while only the Proofreading/Catalog task shows results more in line with the aforementioned hypothesis in the second experimental design. However, when comparing between the Proofreading/Catalog and Prioritizing tasks from these findings, it is clear that, in the first experimental design a greater number of physiological measures show results in line with the aforementioned hypothesis in the latter task than in the former task, whereas in the second experimental design the former task is the only task where a physiological measure shows results in line with the aforementioned hypothesis.

- From the task-wise means of the self-reported measures per intelligibility condition [5.30 - 5.33], we saw that, in each experimental design, there are self-reported measures in all three tasks for which their mean values show a pattern which is in line with the hypothesis that increase in distraction causes increase in stress. However, when comparing among the three tasks from this finding, it is clear that, in the first experimental design a greater number of self-reported measures show results in line with the aforementioned hypothesis in the Presentation task than in the Proofreading/Catalog and Prioritizing tasks each, whereas in the second experimental design the greatest number of self-reported measures that show results in line with the aforementioned hypothesis is from the Prioritizing task, followed by the Presentation task, and followed last by the Proofreading/Catalog task.

It is clear that these latter findings have outlined differences in results among the task types, thus reflecting our hypothesis at least in these outlined cases that the specific type of task does have an effect on the earlier results. It is also clear that the task-wise mean values are most in line with the aforementioned hypothesis in the Proofreading/Catalog task (when compared to the other two tasks) only when these mean values are of the physiological measures in the first experimental design. Thus it is concluded that our hypothesis that these results (i.e, the task-wise means of the physiological measures and the self-reported measures) will be most in line with the corresponding hypothesis for the Proofreading/Catalog task is reflected only in the task-wise mean values of the physiological measures in the first experimental design.

Results might depend on the specific cognitive task type, for example, on the properties of the specific cognitive task type. One such property of cognitive task type is cognitive load (i.e, mental workload, as measured via NASA-TLX) [60], [61], [62]. Further analysis to study this and other such properties with respect to the cognitive tasks used in the study of this thesis and how these properties vary among the tasks might be crucial towards understanding if and/or how the properties might have affected results in this thesis. This analysis might provide additional insights for explaining certain findings from the results in this thesis (and not only the results discussed as part of answering this research question).

6. **Which intelligibility condition has a stronger effect on stress and performance?** We consider all the means and task-wise means computed as part of the results of this thesis together in order to answer this research question.

    - From the means and task-wise means of the physiological measures computed per intelligibility condition, or in other words in this case, when we consider physiological measures as the sole indicator of stress [Tables 5.4 - 5.7, 5.22 - 5.29], we see that, in both the first and the second experimental designs, the silence condition overall causes stress to be the highest. These findings are not in line with our hypothesis that higher stress would be yielded in the high intelligibility condition when compared to the low intelligibility and silence conditions.
    - From the means and task-wise means of the self-reported measures computed per intelligibility condition, or in other words in this case, when we consider self-reported subjective stress ratings as the sole stress indicator [Tables 5.8 - 5.9, 5.30 - 5.33], we see that, in the first experimental design, the silence condition overall causes stress to be the highest. However, in the second experimental design, we see that the low intelligibility condition overall causes stress to be the highest. Again, we see that these

findings are not in line with our hypothesis that higher stress would be yielded in the high intelligibility condition when compared to the low intelligibility and silence conditions.

Thus based on the above findings, considering both hypothesized stress indicators and both experimental designs, it is concluded that overall the silence condition has a stronger effect on stress relative to the high intelligibility and low intelligibility conditions, which is not in line with the proposed hypothesis for this research question.

- From the means of the task performance measures computed per intelligibility condition [Tables 5.10 - 5.11], in the first experimental design, we see that the low intelligibility condition overall causes performance to be the lowest, while in the second experimental design, the high intelligibility condition overall causes performance to be the lowest. This latter finding is in line with our hypothesis that lower performance would be yielded by the high intelligibility condition when compared to the low intelligibility and silence conditions.

From the above finding it is concluded that the low intelligibility condition overall has a stronger effect on performance in the first experimental design when compared to the other two intelligibility conditions, while the high intelligibility condition overall has a stronger effect on performance in the second experimental design when compared to the other two intelligibility conditions. Only the latter conclusion is in line with the proposed hypothesis for this research question.

7. **Are these results consistent across different design variations of the experiment?** We will answer this question based on the results analyzed in each of the previous research questions, as follows:

   - Research question 1: We saw differences in observations between the two experimental designs. Here, the first experimental design showed results more in line with the corresponding hypothesis when considering the indicator of stress as only physiological measures, while the second experimental design showed results more in line with the corresponding hypothesis when considering the indicator of stress as only self-reported subjective stress ratings. These observations are in line with the conclusions (derived from answering the fourth research question) that physiological measures are a more accurate indicator of stress in the first experimental design, while self-reported subjective stress ratings are a more accurate indicator of stress in the second experimental design.

- Research question 2: We saw that results between the experimental designs were more or less the same when considering the indicator of stress as only physiological measures, but when considering the indicator of stress as only self-reported subjective stress ratings, we saw that there were more negative and relatively low magnitude positive correlation values in the second experimental design than in the first experimental design. However, needless to add, the results in both experimental designs conformed to the corresponding hypothesis.
- Research question 3: We see that there is a higher proportion of cognitive tasks whose total score means conformed to the corresponding hypothesis in the second experimental design than in the first experimental design. Moreover, the results in the second experimental design conformed to this hypothesis, while the results in the first experimental design did not.
- Research question 4: We had seen differences in results between the two experimental designs, specifically with regards to the more accurate indicator of stress. However, we also saw, for example, that, in both experimental designs the physiological measures were overall positively correlated with the self-reported measures.
- Research questions 5 and 6: We had seen differences in results between the two experimental designs as discussed earlier when answering these questions. However, we also saw an example where there are similarities between the experimental designs,

while answering the sixth research question: the silence condition causes stress to be the highest in both experimental designs when considering physiological measures to be the sole indicator of stress.

It was hypothesized in this thesis that the second experimental design is methodologically better than the first experimental design and thus will yield results more in line with our research question hypotheses. Here we state a possibly trivial conclusion that we can safely make: the cases referenced in the discussion as part of answering this research question where the results in the second design variation are in line with the corresponding hypotheses whereas the results in the first design variation are not, and the cases referenced in the same where the former are closer to being in line and/or more in line with the corresponding hypotheses in comparison to the results in the first design, all conform to our hypothesis for this research question.

Based on the results and discussion thus far in this thesis, the findings are mixed as to which of the experimental designs conforms more with our hypotheses. There could be multiple ways to go about identifying which experimental design could be the answer to this question. The approach we take in this thesis will consist of identifying the proportion of research questions whose hypotheses were satisfied by each of the experimental designs and then taking the design with the highest proportion as the answer. Thus, taking this approach, we first summarize the following for each research question:

- Research question 1: Each design was closer to being in line with the corresponding hypothesis when considering one of the hypothesized stress indicators, for which the other design would not be in line with the corresponding hypothesis.
- Research question 2: Though results in both experimental designs were in line with the corresponding hypothesis, the results in the second design are potentially more in line with the hypothesis (shown by a greater number of negative and relatively low magnitude positive correlation values in the second design).

- Research question 3: Only results in the second experimental design conformed to the corresponding hypothesis, while results in the first design did not.
- Research question 4: Our hypothesis was reflected in the first experimental design but not in the second design.
- Research question 5: Our hypotheses are reflected overall more in the results of the first experimental design than in the results of the second experimental design.
- Research question 6: Results in the second experimental design conform to the performance aspect of the question hypothesis but not the stress aspect, while results in the first experimental design do not conform to either aspect of the question hypothesis.

Now, combining these findings from all six of the research questions, it is concluded that overall, results in the Between-Groups (randomized) experimental design were more in line with our research question hypotheses than results in the Within-Subjects experimental design. This conclusion is in line with our hypothesis for this research question (research question 7).

7. LIMITATIONS AND FUTURE WORK

7.1 Limitations

- Out of the physiological data collected from the user study, only the EDA and HR physiological signals were analyzed in this research.
- Sample sizes were unequal for EDA and HR in the second experimental design.
- HR and EDA data were not available for two participants in the second experimental design, which may have negatively affected the results.

7.2 Future work

- Variations in demographics of the participants, including age, ethnicity, and level of education, might have been an additional factor in affecting the results. The demographics data collected in the user study of this thesis could be analyzed for verifying this and to study how they might have affected the results. This analysis might provide additional insights for explaining some of the findings from the results.
- As ECG and BVP (Blood Volume Pulse) physiological signals were recorded from participants in both experimental designs in the user study as well, the analysis done in this thesis could be redone including these additional signals.
- Machine learning regression models can be examined whether they are able to predict participants':
  - subjective stress state based on physiological data.
  - performance based on self-reported subjective ratings and physiological data.
- Various types of regression models can be compared in order to determine best fit for the data.

- In this work, the analysis was done with respect to the type of cognitive task, and it was determined whether the task type affected the results. Analysis can be redone taking into account task order as well, and trying to determine how the ordinal rank of a cognitive task during the user study affects the results.

REFERENCES

[1] Sörqvist Patrik. On interpretation and task selection in studies on the effects of noise on cognitive performance. *Frontiers in Psychology*, 5, 2014.

[2] Brian H. Dalton and David G. Behm. Effects of noise and music on human and task performance: A systematic review. *Occupational Ergonomics*, 7(3):143-152, 2007.

[3] M. Haka, A. Haapakangas, J. Keränen, J. Hakala, E. Keskinen, and V. Hongisto. Performance effects and subjective disturbance of speech in acoustically different office types – A laboratory experiment. *Indoor Air*, 19(6):454-467, 2009.

[4] Andrew P. Smith. Noise and aspects of attention. *British Journal of Psychology*, 82(3):313-324, 1991.

[5] Eric Sundstrom, Jerri P. Town, Robert W. Rice, David P. Osborn, and Michael Brill. Office noise, satisfaction, and performance. *Environment and Behavior*, 26(2):195-222, 1994.

[6] Adrian Leaman. Dissatisfaction and office productivity. *Facilities*, 13(2):13-19, 1995.

[7] Robert A. Baron, Mark S. Rea, and Susan G. Daniels. Effects of indoor lighting (illuminance and spectral distribution) on the performance of cognitive tasks and interpersonal behaviors: The potential mediating role of positive affect. *Motivation and Emotion*, 16(1):1-33, 1992.

[8] Jacqueline C. Vischer. The effects of the physical environment on job performance: Towards a theoretical model of workspace stress. *Stress and Health*, 23(3):175-184, 2007.

[9] Jafar Akbari, Habibollah Dehghan, Hiva Azmoon, and Farhad Forouharmajd. Relationship between lighting and noise levels and productivity of the occupants in automotive assembly industry. *Journal of Environmental and Public Health*, 2013, 2013.

[10] Jessica Errett, Erica Eileen Bowden, Marc Choiniere, and Lily M. Wang. Effects of noise on productivity: Does performance decrease over time? *Architectural Engineering - Faculty Publications*, 2006.

[11] Jerry D. Ramsey. Task performance in heat: A review. *Ergonomics*, 38(1):154-165, 1995.

[12] Adrian Leaman and Bill Bordass. Productivity in buildings: The 'killer' variables. *Building Research & Information*, 1(1):4-19, 2010.

[13] Michael W. Eysenck. *Attention And Arousal - Cognition And Performance*. Berlin: Springer-Verlag, 1982.

[14] Staffan Hygge and Igor Knez. Effects of noise, heat and indoor lighting on cognitive performance and self-reported affect. *Journal of Environmental Psychology*, 21(3):291-299, 2001.

[15] M. Spreng. Possible health effects of noise induced cortisol increase. *Noise Health*, 2(7):59-64, 2000.

[16] Thomas Witterseh, David P. Wyon, and Geo Clausen. The effects of moderate heat stress and open-plan office noise distraction on SBS symptoms and on the performance of office work. *Indoor Air*, 14(8):30-40, 2004.

[17] Paul Roelofsen. Performance loss in open-plan offices due to noise by speech. *Journal of Facilities Management*, 6(3):202-211, 2008.

[18] N. Venetjoki, A. Kaarlela-Tuomaala, E. Keskinen, and V. Hongisto. The effect of speech and speech intelligibility on task performance. *Ergonomics*, 49(11):1068-1091, 2006.

[19] Ivana Balazova, Geo Clausen, Jens Holger Rindel, Torben Poulsen, and David P. Wyon. Open-plan office environments: A laboratory experiment to examine the effect of office noise and temperature on human perception, comfort and office work performance. *Indoor Air*, 2008.

[20] Anne Marie Perrotti, Silvana Maria Russo Watson, Stacie I. Ringleb, Anastasia M. Raymer, and Ivan Ash. Effects of noise and audiovisual cues on speech processing in adults with and without ADHD. *International Journal of Audiology*, 53(3), 2014.

[21] Valtteri Hongisto. A model predicting the effect of speech of varying intelligibility on work performance. *Indoor Air*, 15(6):458-468, 2005.

[22] Annu Haapakangas, Miia Haka, Esko Keskinen, and Valtteri Hongisto. Effect of speech intelligibility on task performance - An experimental laboratory study. In *9th International Congress on Noise as a Public Health Problem (ICBEN)*, 2008.

[23] Rainer Thaden, Sabine Schlittmeier, J. Hellbruck, and Michael Vorlaender. The impact of background speech varying in intelligibility: Effects on cognitive performance and perceived disturbance. *Ergonomics*, 51(5):719-736, 2008.

[24] Annu Haapakangas, Valtteri Hongisto, Jukka Hyona, Joonas Kokko, and Jukka Keranen. Effects of unattended speech on performance and subjective distraction: The role of acoustic design in open-plan offices. *Applied Acoustics*, 86:1-16, 2014.

[25] Phil Laether, Diane Beale, and Lucy Sullivan. Noise, psychosocial stress and their interaction in the workplace. *Journal of Environmental Psychology*, 23(2):213–222, 2003.

[26] Michael Pluess. Individual differences in environmental sensitivity. *Child Development Perspectives*, 9(3), 2015.

[27] Samuel Ajayi. Effect of stress on employee performance and job satisfaction: A case study of Nigerian banking industry. *SSRN*, 2018.

[28] Mark A. Staal. *Stress, cognition, and human performance: A literature review and conceptual framework*. National Aeronautics and Space Administration, 2004. https://ntrs.nasa.gov/api/citations/20060017835/downloads/20060017835.pdf

[29] J. C. Wofford and P. S. Daly. A cognitive-affective approach to understanding individual differences in stress propensity and resultant strain. *Journal of Occupational Health Psychology*, 2(2):134–147, 1997.

[30] Frank J. J. M. Steyvers and Anthony W. K. Gaillard. The effects of sleep deprivation and incentives on human performance. *Psychological Research*, 55(1):64–70, 1993.

[31] Siao Z. Bong, Murugappan M., Sazali Yaacob. Methods and approaches on inferring human emotional stress changes through physiological signals: A review. *International Journal of Medical Engineering and Informatics*, 5(2):152-162, 2013.

[32] K. Mohanavelu, R. Lamshe, S. Poonguzhali, K. Adalarasu, and M. Jagannath. Assessment of human fatigue during physical performance using physiological signals: A review. *Biomedical and Pharmacology Journal*, 10(4), 2017.

[33] Inma Mohino-Herranz, Roberto Gil-Pita, Javier Ferreira, Manuel Rosa-Zurera, and Fernando Seoane. Assessment of mental, emotional and physical stress through analysis of physiological signals using smartphones. *Sensors (Basel, Switzerland)*, 15(10):25607-25627, 2015.

[34] Gavin P. Trotman, Jet J. C. S. Veldhuijzen van Zanten, Jack Davies, Clara Moller, Annie T. Ginty, and Sarah E. Williams. Associations between heart rate, perceived heart rate, and anxiety during acute psychological stress. *Anxiety, Stress, & Coping*, 2(6):711-727, 2019.

[35] Mickael Causse, Frederic Dehais, Philippe-Olivier, and Fabrice Cauchard. An analysis of mental workload and psychological stress in pilots during actual flight using heart rate and subjective measurements. In *5th International Conference on Research in Air Transportation (ICRAT 2012)*. Berkeley, 2012.

[36] W. Michael Felts. Relationship between ratings of perceived exertion and exercise-induced decrease in state anxiety. *Perceptual and Motor Skills*, 69(2):368-70, 1989.

[37] Karel A. Brookhuis and Dick de Waard. Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention*, 42(3):898-903, 2010.

[38] Bruce Mehler, Bryan Reimer, Joseph Coughlin, and Jeffery Dusek. The impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, 2138(1):6-12, 2009.

[39] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, and Gerhard Troster. Discriminating stress from cognitive load using a wearable EDA device. In *IEEE Transactions on Information Technology in Biomedicine*, 14(2):410-417, 2010.

[40] David C. Glass, Bruce Reim, and Jerome E. Singer. Behavioral consequences of adaptation to controllable and uncontrollable noise. *Journal of Experimental Social Psychology*, 7(2):244-257, 1971.

[41] Ken I. Hume and Mujthaba Ahtamad. Physiological responses and subjective estimates of soundscape elements: Preliminary results for respiratory rate and EMG responses. In *Proceedings of the International Congress on Noise Control Engineering (Internoise)*. Ottawa, 2009.

[42] Sang Hee Park, Pyoung Jik Lee, and Jeong Ho Jeong. Effects of noise sensitivity on psychophysiological responses to building noise. *Building and Environment*, 136:302-311, 2018.

[43] S. A. Stansfeld. Noise, noise sensitivity and psychiatric disorder: Epidemiological and psychophysiological studies. *Psychological Medicine Monograph Supplement*, 22:1-44, 1992.

[44] Irene van Kamp, R. F. Soames Job, Julie Hatfield, Mary Haines, Rebecca K. Stellato, and Stephen A. Stansfeld. The role of noise sensitivity in the noise-response relation: A comparison of three international airport studies. *The Journal of the Acoustical Society of America*, 116(6), 2004.

[45] Gert Notbohm, Renate Schmook, Sieglinde Schwarze, and Peter Angerer. Patterns of physiological and affective responses to vehicle pass-by noises. *Noise & Health*, 15(66):355-366, 2013.

[46] Helena Jahncke, Valtteri Hongisto, and Petra Virjonen. Cognitive performance during irrelevant speech: Effects of speech intelligibility and office-task characteristics. *Applied Acoustics*, 74(3):307-316, 2013.

[47] Carol L. Mackersie and Natalie Calderon-Moultrie. Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear and Hearing*, 37(p):118S-125S, 2016.

[48] Alexander L. Francis, Megan K. MacPherson, Bharath Chandrasekaran, and Ann M. Alvar. Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort. *Frontiers in Psychology*, 7:263, 2016.

[49] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A Neerincx, and Wessel Kraaij. The SWELL knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*, pp. 291–298, 2014.

[50] Charles D. Spielberger, R. L. Gorsuch, R. Lushene, P. R. Vagg, and G. A. Jacobs. *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press, 1983.

[51] Charles D. Spielberger, R. L. Gorsuch, R. Lushene, P. R. Vagg, and G. A. Jacobs. *State-Trait Anxiety Inventory for Adults*. Mind Garden, Inc, 1983. https://oml.eular.org/sysModules/obxOML/docs/id_150/State-Trait-Anxiety-Inventory.pdf

[52] Lewis R. Goldberg. A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7:7-28, 1999.

[53] Christopher J. Soto and Joshua J. Jackson. Five-factor model of personality. In Dana S. Dunn (Eds.), *Oxford Bibliographies in Psychology*. Oxford University Press, 2020.

[54] Hiske Calsbeek, Mieke Rijken, Henegouwen, Gerard P. van Berge Henegouwen, and Joost Dekker. Factor structure of the Coping Inventory for Stressful Situations (CISS-21) in adolescents and young adults with chronic digestive disorders. *The Social Position of Adolescents and Young Adults with Chronic Digestive Disorders*, pp. 83-103. Utrecht:NIVEL, 2003.

[55] Yoonmi Choi, Eunsoo Moon, Je Min Park, Byung Dae Lee, Young Min Lee, Hee Jeong Jeong, and Young In Chung. Psychometric properties of the Coping Inventory for Stressful Situations in Korean adults. *Psychiatry Investigation*, 14(4):427–433, 2017.

[56] N. D. Weinstein. Individual differences in reactions to noise: A longitudinal study in a college dormitory. *Journal of Applied Psychology*, 63(4):458-466, 1978.

[57] *E4 wristband*. Empatica. https://www.empatica.com/research/e4/. Accessed 19 August 2021.

[58] *Actiwave Cardio*. CamNtech. https://www.camntech.com/cardio/. Accessed 25 September 2021.

[59] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139-183, 1983.

[60] Frederick Reif F. *Applying Cognitive Science to Education: Thinking and Learning in Scientific and Other Complex Domains*. Cambridge, MA: The MIT Press, 2010.

[61] Lauren R. Kennedy-Metz, Hill L. Wolfe, Roger D. Dias, Steven J. Yule, and Marco A. Zenati. Surgery task load index in cardiac surgery: Measuring cognitive load among teams. *Surgical Innovation*, 27(6):602–607, 2020.

[62] David Harris, Mark Wilson, and Samuel Vine. Development and validation of a simulation workload measure: The simulation task load index (SIM-TLX). *Virtual Reality*, 24(4):557-566, 2020.