SEX DIFFERENCES IN RESPONSE TO A LARGE 200-PERSON AUDIENCE USING THE TRIER SOCIAL STRESS TEST IN PRE-RECORDED VIRTUAL REALITY

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Garrett Shawn Byron

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Psychology

September 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Sex Differences in Response to a Large 200-Person Audience using the Trier Social Stress Test in Pre-Recorded Virtual Reality

By

Garrett Shawn Byron

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Clayton Hilmert
Chair

Dr. Erin Conwell

Dr. Jeffrey Johnson

Dr. Britt Heidinger

Approved:

October 6th, 2022
Date

Dr. Mark Nawrot
Department Chair

ABSTRACT

The Trier Social Stress Test (TSST) induces stress in the lab by having participants complete challenging tasks in front of an audience. The TSST has been adapted to virtual reality (VR), eliminating in-person audience variability that could explain sex differences in stress responses. Furthermore, the VR-TSST facilitates the examination of factors difficult to investigate in person, such as the effect of a large, 200-person audience. This study compared male and female physiological and psychological responses to an in-person 2-person TSST and a prerecorded VRTSST with audience sizes of 2 and 200 persons. Results indicated that only males had statistically significant cortisol reactivity to the TSST and responded with more positive affect, arousal, and lower ratings of stress than females. In the VR conditions these differences were less apparent, suggesting that the sex differences may be a result, at least in part, of in-person audience variability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. OVERVIEW

Psychological stress research has increased our understanding of how stressful situations prompt psychophysiological reactivity and influence the development of adverse health outcomes like cardiovascular diseases, and diabetes (Garfin, Thompson, & Holman, 2018; McEwen, 1998). Acute and chronic stressors contribute to the development of these diseases through the dysregulation of bodily systems that respond to stress, for example, the hypothalamic-pituitary-adrenal (HPA) axis (Burford, Webster, & Cruz-Topete, 2017; Hamer, Endrighi, Venuraju, Lahiri, & Steptoe, 2012). Indeed, psychophysiological research has found that high levels of cortisol (i.e., a measure of HPA axis activation) prompted by chronic stressors (i.e., financial and work stress) (Kivimäki et al., 2006; Moran, Ommerborn, Blackshear, Sims, & Clark, 2019) are significantly associated with higher risk of heart disease (Iob & Steptoe, 2019), the leading cause of death in the United States (Virani et al., 2021).

While heart disease affects both males and females, there are important differences in male and female experiences of this stress-induced disease. Prior to menopause, females are less likely to develop heart disease than similarly aged males (Maas & Appelman, 2010). This sex difference may be due, at least in part, to differences in male and female psychophysiological responses to stress. That is, psychophysiological reactions to stress such as increases in cortisol, blood pressure, heart rate, and anxiety may be more extreme in males than females experiencing the same stressor. This may lead to differences in symptoms and health outcomes.

To better understand the relationship between stress responses and health outcomes like heart disease, laboratory research has used stress induction paradigms to examine physiological and psychological responses to stress. One such paradigm, the Trier social stress test (TSST),

elicits reliable psychophysiological stress responses, including increases in the stress hormone cortisol, blood pressure, negative affect, perceptions of stress, and self-reported anxiety in participants (Goodman, Janson, & Wolf, 2017; Hellhammer & Schubert, 2012; Kudielka, Hellhammer, & Kirschbaum, 2007).

The TSST was first reported by Kirschbaum and colleagues (1993). More recently, virtual reality Trier social stress test (VR-TSST) adaptations have been used to eliminate audience response confounds and afford opportunities to manipulate variables that would be difficult to manipulate in-person (IP) – such as audience size (Mostajeran, Balci, Steinicke, Kühn, & Gallinat, 2020). Although IP-TSST and VR-TSST protocols have been shown to elicit robust responses in cortisol, blood pressure, heart rate, and increases in negative affect (Helminen, Morton, Wang, & Felver, 2021; Santl et al., 2019; Shiban et al., 2016; Zimmer, Buttlar, Halbeisen, Walther, & Domes, 2019), sex differences in responses to the TSST remain unclear.

One meta-analysis reported sex differences in cortisol responses to the IP-TSST with males having higher salivary cortisol at peak and recovery periods than females (J. Liu et al., 2017). However, other IP-TSST studies have reported no sex differences in cortisol responses (M. M. Kelly, Tyrka, Anderson, Price, & Carpenter, 2008). It may be that IP audiences sometimes vary in their responses to male and female participants. The use of a VR-TSST paradigm controls for such variation.

Although a VR-TSST controls for variations in audience behavior within a study, significant variability exists in the presentation of the VR-TSST between studies. This variability likely contributes to inconsistencies in sex differences across studies using VR-TSST protocols (Helminen, Morton, Wang, & Felver, 2019; Santl et al., 2019). For instance, the use

of different technologies in previous studies has led to variations in the appearance and realism of animated VR-TSST environments and audience avatars. Indeed, a meta-analysis suggested that more "immersive" or realistic VR-TSST technologies elicited greater physiological responses (Helminen et al., 2019). The present study optimized the realism of the VR-TSST by using an immersive recording of a real (non-avatar) audience.

Little is known about how participants may respond differently to audiences that differ significantly in size. While there is some indication that a larger audience may elicit stronger psychophysiological responses (Bosch et al., 2009), this research is limited by overall small audience sizes (1 vs. 4 audience members). This is likely due to the resource requirements needed to run a large-audience study in-person. Also, research has not considered sex differences in response to different audience sizes during the TSST.

This study examined sex differences in psychophysiological responses to a classic TSST protocol and two VR-TSST protocols with variations in audience size (i.e., an in-person 2-person audience, a VR 2-person audience, and a VR 200-person audience). This study also employed a novel VR-TSST using an immersive, 360-degree, pre-recorded real audience in a real auditorium to emulate the physical environment of the in-person TSST.

The following literature review begins with a brief history of stress and health research, including the origin of the TSST. Following this, literature validating the use of VR for TSST research is considered. Then reviews of IP and VR-TSST research concerning audience size and sex differences are presented.

1.1. Literature Review

The psychophysiological study of stress has developed quickly over the course of four decades since the coining of the term "stress", by Hans Selye in the mid-20th century (Selye,

1936). The use of "stress" as a psychophysiological measure was motivated by Selye's finding that stressed rats develop adrenal hyperactivity and peptic ulcers, demonstrating that psychosocial stress could result in physiological health problems (Selye, 1936; Tan & Yip, 2018). As psychosocial stress research progressed researchers began to look at specific stress-related health outcomes finding that multiple acute instances of stress resulted in hyperphagia and obesity in rats (Rowland & Antelman, 1976). Further research in the field allowed connections to be drawn between stress and health in humans. For instance, research suggested that failure to adjust to psychosocial stress resulted in a broad range of health disturbances (i.e., depression, heart syndrome) (Aakster, 1974).

To identify and understand individual differences and situational factors that moderate stress responses in humans it became necessary to find a reliable method of eliciting stress. Early paradigms for eliciting stress in participants; namely, the cold pressor task (Lamotte, Boes, Low, & Coon, 2021), the Stroop task (Stroop, 1935), and various public speaking tasks were unable to reliably elicit physiological responses to stress (e.g., cortisol, prolactin, adrenocorticotropic hormone, blood pressure, and heart rate increases). In 1993 Kirschbaum and colleagues published a method of reliably eliciting psychophysiological stress responses using a paradigm that evokes social evaluative threat in participants (Sally S. Dickerson, 2008).

1.1.1. The Trier Social Stress Test (TSST)

The TSST involves two stress elicitation strategies such that participants are asked to complete an impromptu public speech and a mental arithmetic task in front of an evaluative confederate audience. The effects of stress can be seen during the TSST by comparing physiological parameters during a resting baseline period to those parameters during anticipatory, task, and recovery periods (Kirschbaum et al., 1993). This paradigm elicits reliable

psychophysiological stress responses, including increases in the stress hormone cortisol, blood pressure, negative affect, perceptions of stress, and self-reported anxiety (Goodman et al., 2017; Hellhammer & Schubert, 2012; Kudielka et al., 2007). The TSST has become the "gold standard" for producing robust physiological responses in the lab, typifying laboratory stress reactivity research for more than 25 years.

One critical component of the TSST is the presence of an audience. That is, audience presence has been found to directly influence the extent to which participants experience psychological and physiological stress. For example, one study directly compared a classic IPTSST to a placebo TSST in which participants performed a speech and math task without the presence of an experimenter or an audience. Participants in the placebo TSST condition had no cortisol response and self-reported less stress than participants who completed the classic in-person, audience present TSST (Het, Rohleder, Schoofs, Kirschbaum, & Wolf, 2009).

The traditional, IP-TSST is not without potential pitfalls. Variability in audience behavior and subtle differences in protocols have been shown to significantly influence participant responses. For instance, in one study, participants performing a TSST in the presence of a supportive or "friendly" audience had no cortisol response and no increase in negative affect (Wiemers, Schoofs, & Wolf, 2013). Furthermore, the presence or absence of an experimenter during the TSST reverses the impact of the audience on participants' blood pressure reactivity (Hilmert, Kulik, & Christenfeld, 2002). These sometimes-subtle differences in audience behavior and TSST protocols can be controlled with a standardized virtual reality TSST protocol. Furthermore, the use of a pre-recorded, virtually presented audience would allow for resource intensive variables like audience size to be examined.

1.1.2. The Virtual Reality TSST

The development of virtual reality (VR) technology, specifically head mounted displays, has allowed for the presentation of environments in an immersive manner that replaces the audio and visual input of the immediate physical environment. These environments may be static images, live-stream videos, pre-recorded videos, as well as animated environments with which a user may interact. This technology has been professionally developed for medical training (Seymour et al., 2002), the treatment of post-traumatic stress disorder (Gerardi, Cukor, Difede, Rizzo, & Rothbaum, 2010; Gonçalves, Pedrozo, Coutinho, Figueira, & Ventura, 2012; Rothbaum et al., 1999), and other anxiety disorders (Powers & Emmelkamp, 2008). VR technology provides a new medium in which to conduct psychophysiology research and has led to the development of virtual reality versions of the Trier social stress test (VR-TSST).

The VR-TSST follows the same protocol as the classic, in-person (IP) TSST. The primary difference is that the audience is presented virtually rather than physically present. One benefit of this is that audience responses are pre-recorded or pre-programmed (Fallon, Careaga, Sbarra, & O'Connor, 2016; Jönsson et al., 2010; Montero-López et al., 2016; Ruiz et al., 2010). Therefore, a VR-TSST provides the same standardized audience responses to all participants regardless of their sex or the content or quality of the performances.

Early studies that have investigated the use of VR-TSST protocols reported increases in cortisol (O. Kelly, Matheson, Martinez, Merali, & Anisman, 2007), heart rate (Jönsson et al., 2010; Westenberg et al., 2009) and blood pressure (O. Kelly et al., 2007; Kotlyar et al., 2008). A metaanalysis of thirteen VR-TSST studies used within-subjects effect sizes to conclude that the effect of VR-TSST protocols on cortisol production was “moderate” (Helminen et al.,

2019). In addition, a comparison of VR-TSST protocols using various presentation methods, (e.g., head mounted displays, 2D screens, 3D screens, or a projection on the surrounding walls) concluded that there was a greater cortisol response to VR-TSST presentations that replaced more sensory inputs (i.e., had greater immersivity) (Helminen et al., 2019).

Studies have directly compared IP and VR-TSST effects. One study found that participant cortisol responses to a VR-TSST protocol were comparable to an identical IP-TSST protocol (Zimmer et al., 2019). Another found that although both protocols elicited similar self-reported stress appraisals, the IP-TSST resulted in greater cortisol responses compared to its VR counterpart (Shiban et al., 2016). More recently a meta-analysis of 16 comparison studies concluded that the VR-TSST is as effective as an IP-TSST at eliciting physiological reactivity and self-reported stress responses (Helminen et al., 2021). Therefore, with this new technology researchers can explore the effects of stress-related variables that would be difficult to manipulate with an IP protocol. Here, audience size was considered.

1.1.3. Audience Size In-Person

The presence of an audience promotes the threat of social evaluation, exacerbating psychophysiological responses to TSST paradigms (Het et al., 2009). This suggests that variations in audience size may prove influential. A typical TSST paradigm employs the use of a small IP audience of two to five members (Helminen et al., 2019; Kudielka & Kirschbaum, 2005). Few studies have examined the psychophysiological effects of audience size due to the resource requirements involved in having a large in-person confederate audience trained to react identically for each participant. However, a small number of studies have explored the effects of IP audience sizes in various paradigms.

Previous work has generally concluded that audience size is positively associated with participants' self-reported anxiety (Knowles, 1983) and feelings of being socially judged (Latané & Harkins, 1976). In a study of actors' anxiety on stage in response to different audience sizes, Lemasson (2018) found that performing a play in front of a large audience of 128 members resulted in greater self-reported anxiety compared to smaller audiences of 30 members and 8 members. A study of audience size and speech performance found that an audience size of four members elicited greater heart rate, pre-ejection period, and cortisol than an audience of one member (Bosch et al., 2009). Although this handful of studies suggests that a larger audience will elicit stronger psychophysiological stress responses, these studies are limited by using a specialized sample of participants (i.e., actors) and by utilizing a relatively small, large audience (i.e., 4 members).

1.1.4. Audience Size in Virtual Reality

The use of VR technology in TSST research allows for easy manipulation of the resource intensive variable audience size. However, few studies have done so. One study of virtual audience size and social anxiety used an audience of virtual humans (e.g., animated humans) to induce anxiety in participants during a public speaking task (Mostajeran et al., 2020). Participants were asked to give a speech in front of 3, 6, and 15 virtual humans presented through a head mounted display. Interestingly, the small audience of three virtual humans evoked significantly higher HR than the larger audiences of 6 and 15 virtual humans. This small variation in virtual audience size resulted in small differences in physiological responses. Mostajeran et al. (2020) concluded that an audience of 3 virtual humans successfully induced social anxiety but that the introduction of additional virtual humans did not elicit further social anxiety.

Alternatively, the counter-intuitive results of the Mostajeran et al. (2020) study may have been in part, due to the use of animated avatar, or "virtual human" audiences. A study of participant perceptions of virtual humans in virtual reality found that participants were emotionally affected by the attitudes (e.g., friendly, unfriendly, sad) of photorealistic virtual humans more than they were by more simply animated virtual humans (Zibrek, Martin, & McDonnell, 2019). Furthermore, participants reported a preference for more realistic virtual humans. Thus, it is possible that a larger number of animated avatar audience members increases the unrealistic nature of the virtual situation, thereby decreasing the fear of social evaluation essential to the effectiveness of the TSST (S. S. Dickerson, Mycek, & Zaldivar, 2008). To mitigate this possibility, VR technology now allows us to create 360-degree, 3D video recordings of real-world stimuli that can be presented as a remote virtual environment. This study used this new technology to examine the effects of audience size on stress responses in males and females.

1.1.5. Sex Differences In-Person

Although the TSST has been validated and used to examine psychophysiological stress responses in participants, it is unclear if sex differences exist. One meta-analysis of 34 IP-TSST studies looked at salivary cortisol reactivity across the study session. Overall, the analysis showed no sex differences in salivary cortisol at baseline, but that males had higher peak cortisol levels immediately following the TSST and after post task recovery periods compared to females (J. Liu et al., 2017). Other studies have found no sex differences in cortisol responses even when controlling for hormone variations (i.e., oral contraceptives, birth control, menstrual cycles) (M. M. Kelly et al., 2008; J. Liu et al., 2017). When psychological

differences are explored, females self-report higher negative affect, fear, and stress to the TSST (M. M. Kelly, Forsyth, & Karekla, 2006; M. M. Kelly et al., 2008).

Variations in audience behavior specific to the sex of the participant may affect sex differences. While an IP audience is asked to respond in the same manner to all participants it is possible that audience behaviors differ with the sex of the participant. For example, male speakers tend to be rated as more knowledgeable and convincing than female speakers (Aalberg & Jenssen, 2007) regardless of rater's gender (Boring, 2017). Such differences in perceptions could potentially lead to subtle differences in audience responses differentially impacting male and female stress reactivity (Goodman et al., 2017). The use of a more controlled VR-TSST would allow researchers to keep all features of audience behavior consistent for each participant regardless of their sex, gender, or performance during the TSST, thereby improving the consistent delivery of social evaluative threat.

1.1.6. Sex Differences in Virtual Reality

Two studies have looked at sex differences in physiological and psychological measures using VR-TSST paradigms. One study reported primarily mixed results involving self-reported stress, heart rate, heart rate variability (i.e., an index of parasympathetic influence on the heart), and electro-dermal activity (i.e., an index of sympathetic activity). In comparison to females, males had generally lower heart rate and higher heart rate variability, higher electrodermal activity prior to the task, and similar self-reported stress responses (Q. Liu & Zhang, 2020). A second study of gender differences in response to a VR-TSST found no differences in heart rate, cortisol, or electrodermal activity, but that only self-reported stress significantly differed (Santl et al., 2019). Specifically, Santl et al., reported that females reported higher stress than

males. A more comprehensive comparison of male and female responses to a VR-TSST is clearly necessary.

1.2. Hypotheses

The current study tested the following hypotheses:

1) Males would have greater physiological stress reactivity (cortisol, SBP, DBP, HR) than females overall.
2) Females would have greater psychological stress responses than males overall.
3) A VR 200-person audience would elicit greater psychophysiological reactivity than a VR 2-person audience.
4) Sex differences in psychophysiological responses to stress would be different using VR, in which audience responses are perfectly standardized, compared to an IP audience.

Previous research has suggested that although effects of the VR-TSST are somewhat smaller than those of a traditional IP-TSST, the VR-TSST is as effective as an in-person TSST at eliciting stress responses (Helminen et al., 2021). I expect to replicate this finding. However, it is not clear if the effect of a VR 200-person audience will be more similar to an IP 2-person audience or if it will elicit greater psychophysiological stress. Because little is known about audience size effects in general, this is considered an exploratory issue.

2. METHODS

2.1. Participants

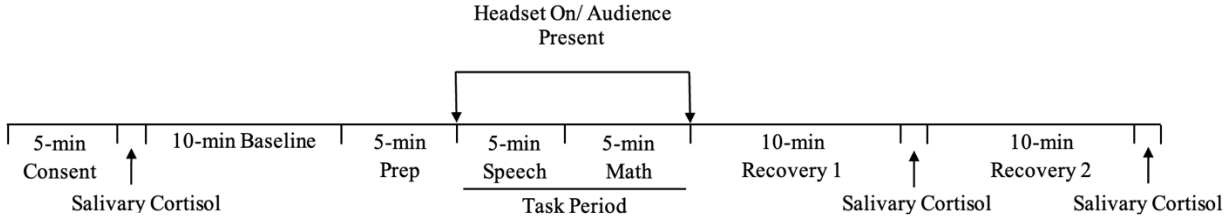
A total of 201 North Dakota State University students participated in this study (61.69% female). Participants were recruited through the psychology department's student subject pool which offers course credit for participation. Saliva was collected in a subset of participants (n = 159) for cortisol assay. Cardiovascular measures were recorded continuously for a subset of participants (n = 111). All procedures were approved by the Institutional Review Board of North Dakota State University.

2.2. Procedure

Participants scheduled a 90-minute lab appointment between 11am-5pm in order to mitigate the effects of diurnal rhythm on salivary cortisol (S. S. Dickerson & Kemeny, 2004). Participants were randomly assigned to one of 3 audience conditions: an in-person 2-person audience (IP 2), a VR 2-person audience (VR 2), or a VR 200-person audience (VR 200). The experiment timeline is shown in figure 1.

Figure 1

Laboratory Experiment Timeline



Note. "Prep" = Preparation.

On arrival to the lab session participants provided informed consent. During the session participants were asked to complete a five-minute speech and a five-minute math task in one of the three confederate audience conditions. During the lab session participant physiological

stress response measures were recorded, including systolic blood pressure (SBP), diastolic blood pressure (DBP), heart rate (HR), and salivary cortisol.

After consent, a Salivette was used to take a baseline salivary cortisol sample. That is, the participant was asked to chew on a small cotton dental roll (Salivette) and then hold it under their tongue for three minutes. The participant was then fitted with a Finometer finger and arm cuff to record cardiovascular parameters. Following this set-up period participants were instructed to relax for 10-minutes to establish baseline readings and orient themselves in the study environment. After baseline measures were completed, participants were asked to prepare a 5-minute speech in which they explained to hiring managers why they are the best person for a job. They were next informed that their performance would be recorded and analyzed by experts in speech articulation and body language. The following procedure varied by condition.

Participants in the *IP 2 condition* were informed that a 2-person audience would soon arrive. The experimenter waited outside the study room for 2-minutes and then welcomed a 2person confederate audience and asked them to be seated across from the participant. Prior to arriving, the confederate audience had been instructed to respond in a non-positive manner to participant speeches (see *Audience Stimuli* below). The participant was then instructed to begin their speech. Following the 5-minute speech task participants were instructed to perform oral arithmetic, counting backwards from 2083 by 13's as quickly and accurately as possible. During the arithmetic, the experimenter pointed out errors and reminded the participant that it is important to perform the task quickly and accurately. On completion of the second 5-minute task the audience members were asked to leave, and the participant was instructed to sit for a 10-minute Recovery 1 period.

Participants assigned to either the *VR 2 or VR 200 conditions* were fitted with an Oculus headset after being informed that their performance would be recorded. The Oculus headset provided an immersive virtual reality environment. A 2-minute orientation period was given to participants while they viewed an empty auditorium where the speech was to be given. Following this period of initial immersion, the experimenter explained that they were switching to "the audience feed" in the same auditorium (see details below). The participant next completed the 5minute speech task and the oral arithmetic math task (see above). After the completion of the math task the Oculus headset was removed, and the participant was instructed to sit for a 10-minute Recovery 1 period.

Across all conditions, following the completion of the TSST, the participant was asked to provide information regarding their subjective experience using the Stress and Arousal Checklist (SACL) and Positive and Negative Affect Scale (PANAS) during the 10-minute Recovery 1 period. Following this first 10-minute recovery period the participant provided a second salivary cortisol sample. Subjects then sat for another 10-minute Recovery 2 period after which a final third salivary cortisol sample was taken, and the subject was debriefed.

2.3. Physiological Measures

2.3.1. Salivary Cortisol

Saliva samples were collected using Salivettes (Germany), a small cotton dental roll held under the tongue for three minutes. Three Salivettes were collected during the laboratory session, first following consent, and then at 20 minutes and 30-35 minutes after the initiation of the speech task (see Figure 1). According to recommendations made in the literature, this provided baseline, peak, and early recovery levels of cortisol respectively (S. S. Dickerson &

Kemeny, 2004). Saliva samples were shipped to Salimetrics (CA) for cortisol assays in duplicate.

2.3.2. Cardiovascular Functioning

A Finometer Pro® (Finapres, Netherlands) was used to record continuous SBP, DBP, and HR for the duration of the lab session. The Finometer Pro® recorded beat-to-beat cardiovascular measurements through a pressurized finger cuff placed on the middle finger of the non-dominant hand (Jansen et al., 2001; Schutte, Huisman, van Rooyen, Oosthuizen, & Jerling, 2003). To calculate baseline SBP, DBP, and HR, the last five minutes of the baseline period values were averaged. All other period values (Preparation, Speech Task, Math Task, Recovery 1, and Recovery 2) were calculated by averaging blood pressure and heart rate values over the entire period.

2.4. Psychosocial Measures

Demographic and psychosocial measures were completed by participants via a tablet computer utilizing NDSU's Qualtrics online survey host as a secure method of assessment following the completion of the Speech and Math tasks.

2.4.1. Stress and Arousal Checklist (SACL)

To measure the subjective experience of stress and arousal during the tasks, participants completed the SACL following the end of the task period (Mackay, Cox, Burrows, & Lazzerini, 1978). Participants were asked to reflect on the extent to which they experienced 20 stress- and arousal-related emotions "during the tasks" using a 4-item scale ranging from "definitely no" to "definitely yes" (see Appendix). Higher values on the scales indicated more stress and more arousal. In the current study the SACL stress sub-scale items met reliability criteria, Cronbach's $\alpha = 0.91$, and were averaged to create the SACL stress index. The SACL arousal sub-scale

items did not meet acceptable reliability, Cronbach's $\alpha = 0.52$. Thus, results involving the SACL arousal index should be interpreted with caution.

2.4.2. Positive and Negative Affect Schedule (PANAS)

The Positive and Negative Affect Schedule – Expanded form (PANAS-X) (Watson & Clark, 1994), asked participants to rate the extent to which they felt 20 general positive and negative emotions "during the task," on a scale from 1 "not at all" to 5 "extremely" (see Appendix). Higher values on the subscales indicated more positive and more negative affect. In the current study, both the PANAS general positive affect sub-scale items and the PANAS general negative affect sub-scale items met reliability criteria, Cronbach's $\alpha = 0.87$.

2.4.3. Perceptions of Audience

In consideration of how VR and audience size affected perceptions of an audience, six questions asked participants to rate how attentive, cheerful, supportive, stressful, judgmental, and sleepy the audience seemed. An additional question also asked how well the audience could hear the participant to see if participants believed this deception. These questions were answered on a 5-point Likert scale with 1 = "not at all" and 5 = "extremely."

2.4.4. Effort and Engagement

Seven questions asked participants how much effort and energy they put into the task, how hard they tried, and how engaged, comfortable, involved, and confident they were. Participants responded on 9-point Likert scales. In the current study the seven effort and engagement questions met reliability criteria, Cronbach's $\alpha = 0.83$. The items were combined to create an Effort and Engagement index in which higher ratings indicated more effort and more engagement.

2.5. Audience Stimuli

2.5.1. In-Person Audience

For the IP 2 audience condition, two undergraduate confederates dressed in lab coats and carried clipboards. These audience members were trained to appear evaluative (i.e., periodically taking notes) and disinterested during a participant's performance of the speech and math tasks. Due to their proximity to the participant, confederates were trained to make subtle actions (e.g., looking away from the participant, glancing at a watch, or slightly shaking their head) to imply disapproval. Two-person audiences were made up of two females or a male and a female.

2.5.2. Virtual Reality (VR) Audiences

The immersive prerecorded VR audience stimuli were created using a Vuze (Human Eyes, New York) 360° 3D camera in front of a 300-person theater. The camera was situated behind a table to emulate the table and height of participants seated in the lab. Before videotaping, the audience was given instructions to appear disinterested and to display behaviors such as looking at a watch or staring off into space approximately every 20 seconds during task periods. A timer was visible to the audience to coordinate actions during the 12-minute recording (1.25-minute pre task, 5-minute speech, .75-minute interim, 5-minute math task). During the recordings the room was silent except for the rustling of the audience. The ambient sound was included in the video playback.

To increase the immersive nature of the VR experience there was a coordinated "interaction" between the experimenter and VR audience prior to the beginning of the speech. Specifically, at forty-five seconds into the recording playback, the experimenter asked the audience to wave if they could hear them. Then, as if in response, the prerecorded audience

waved to the camera. To mirror the in-person lab experience a confederate experimenter in a white lab coat could be seen standing to the left of the participant and appearing to take notes throughout the recording.

2.5.2.1. VR 2-Person Audience

The VR 2 audience wore lab coats and held clipboards that they occasionally took notes on. They were seated in the center of the front row in the theater so that it appeared that the audience was the same distance from the participant as the IP 2 audience. Equivalent to the IP 2 audience, the VR 2 audience was made up of two female confederates or a male and female confederate trained to respond with disinterested and evaluative actions (see above).

2.5.2.2. VR 200-Person Audience

Undergraduate students were recruited to act as the 200-person audience by offering class credit for participation. Instructions with a list of suggested behaviors were distributed to the audience prior to recording. VR 200 audience members were told to act as if they were listening to "a bad lecture in a class they don't like," and to make the suggested behaviors subtle so that they did not seem disingenuous. Two members of the audience left the theater at two points during the participant's performance (at three minutes and eight minutes) to enhance the stressful nature of the situation.

2.6. Analyses

Primary analyses of physiological reactivity included mixed-model ANOVAs with a physiological measure as the repeated factor (cortisol, SBP, DBP, HR) and condition (IP 2, VR 2, and VR 200) and sex (male, female) as between-subject factors. Significant effects were explored with post-hoc least significant difference (LSD) analyses. In the case of sex and audience condition interactions, mixed-model analyses considered male and female patterns of

physiological reactivity across conditions separately. Then, to address the hypothesis that sex differences may differ in VR compared to IP 2, the VR 200 condition and VR 2 condition were collapsed into a single VR condition for subsequent mixed-model ANOVA analyses. Because participants who provided saliva tended to have higher average SBP (p = .053) primary analyses of SBP included the covariate "saliva provision" (0 = no; 1 = yes). Analysis of psychological responses included univariate ANOVAs comparing responses between three conditions and two sexes.

3. RESULTS

3.1. Participants

The average age of participants was 19.29 years (SD = 2.56). The race and ethnicity of the sample was representative of the surrounding community with 85.07% “White/Caucasian,” 6.97% “Asian,” 6.47% “Black or African American,” 2.99% “Hispanic or Latino,” and 3.5% of other ethnicity/cultural background.

3.2. Cortisol

A natural log (x + 1) transformation was applied to reduce skew in cortisol data. Cortisol values that were more than three standard deviations from the mean were excluded from analyses (n = 6; 4 males). The reported means and standard deviations are computed from the untransformed data. The primary 2 (sex) by 3 (audience condition) by 3 (cortisol timepoint) mixed-model ANOVA indicated that there was a significant within-subjects effect of timepoint (Baseline, Recovery 1, and Recovery 2) on cortisol concentrations (Table 1), $F(2,268) = 50.91$, $p < .001$, $\eta^2 = .28$. Pairwise LSD comparisons indicated that overall, the TSST protocol elicited a significant increase (reactivity) followed by a significant decrease (recovery) in cortisol concentrations regardless of audience condition, (Table 1), $F(4,268) = 1.26$, $p > .20$.

A significant sex by timepoint effect, $F(2,268) = 27.26$, $p < .001$, $\eta^2 = .17$, was explored with LSD comparisons. These analyses indicated that male and female cortisol concentrations at baseline did not statistically differ (M = .21 and .21 μg/dL; SD = .12 and .12, respectively), $p > .90$. However, males had significantly higher concentrations than females at Recovery 1 (M = .45 and .25 μg/dL; SD = .20 and .21, respectively) and Recovery 2 (M = .34 and .21 μg/dL; SD = .16 and .20, respectively), ps < .05. To further examine the sex by timepoint effect for reactivity LSD comparisons showed that, for males, average Baseline cortisol was significantly

lower than average Recovery 1 cortisol ($p < .001$), suggesting that males had cortisol reactivity to the TSST.

Table 1

Physiological Parameters by Study Period

Physiology	Baseline	Preparation	Task*	Recovery 1	Recovery 2
Cortisol (μg/dl)	0.20_a (0.12)	-	-	0.31_b (0.22)	0.25_c (0.17)
SBP (mmHg)	121.42_a (14.49)	140.20_b (19.83)	155.25_c (22.74)	141.82_b (21.21)	134.78_d (18.04)
DBP (mmHg)	73.27_a (9.63)	83.04_b (11.76)	93.92_c (14.32)	85.95_b (12.63)	82.20_d (11.58)
HR (bpm)	81.34_a (12.29)	87.50_b (13.42)	sp: 94.30_c (14.31) / ma: 91.55_d (14.65)	84.06_e (13.00)	80.83_a (11.15)

Note. Mean physiological parameter by study period. The values in parentheses represent standard deviation (SD) values.
[a]SBP = Systolic Blood Pressure; DBP = Diastolic Blood Pressure; HR = Heart Rate. [b]Task represents an average of speech and math task values except for HR, where values are reported for (sp)eech / (ma)th tasks separately.
*Different subscripts within rows denote statistically significant mean differences, ps < .05.

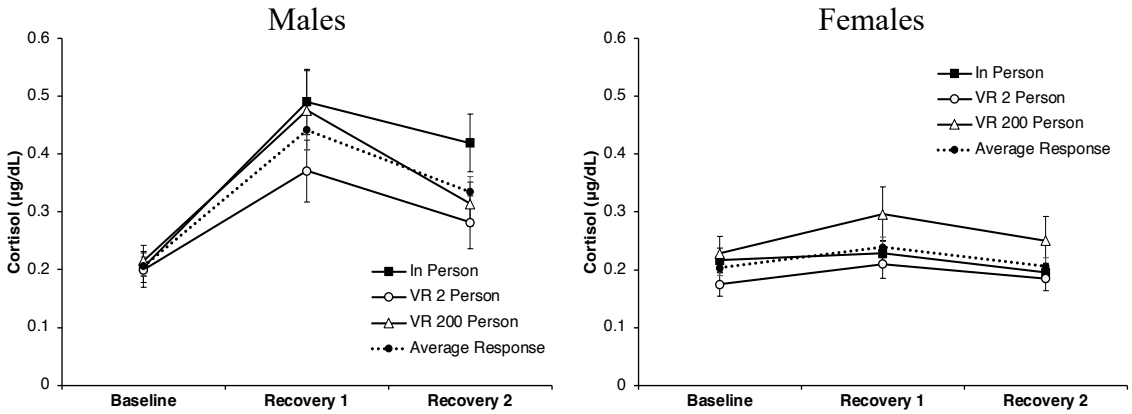
Recovery 2 average cortisol was intermediate to, and significantly different from, Baseline and Recovery 1 average cortisol concentrations ($p < .05$). For females, there was a marginally significant increase in cortisol from Baseline to Recovery 1 ($p = .06$). Females also had a statistically significant decrease in average cortisol from Recovery 1 to Recovery 2 ($p < .001$). The condition by sex by cortisol timepoint three-way interaction was not statistically significant ($p > .05$).

Consistent with the patterns revealed above, between subjects effects showed that average cortisol was higher in males (M = .33 μg/dL, SD = .14) than females (M = .22 μg/dL, SD = .14), $F(1,134) = 21.36$, $p < .001$, $\eta^2 = .14$. Also, there was a marginally significant between-subjects effect of condition on average cortisol concentration, $F(2,134) = 2.90$, $p = .058$, $\eta^2 = .04$, suggesting that conditions differed in the average amount of cortisol elicited (figure 2). Follow-up pairwise least significant difference (LSD) comparisons revealed that

average cortisol was significantly lower in the VR 2 condition (M = .24 μg/dL; SD = .14) than both the VR 200 and IP 2 conditions (M = .29 and .30 μg/dL; SD = .14 and .14, respectively), *p*s < .05. Average overall cortisol concentrations did not statistically differ in the IP 2 and VR 200 conditions, $p > .05$.

Figure 2

Mean Salivary Cortisol by Sex and Condition



Note. Cortisol patterns for males (left), and cortisol patterns for females (right) over the course of the Trier Social Stress Test by condition. Error bars represent the standard error of the mean.

A supplementary analysis included the use of oral contraceptives as a covariate in the primary analysis (Gervasio, Zheng, Skrotzki, & Pachete, 2021). The inclusion of this variable did not explain a significant amount of variance ($p > .30$) or change the reported results.

3.2.1. Cortisol Correlations with Psychological Responses

Correlations between the cortisol measures and psychological responses to the TSST protocols were computed to explore potential mediators of the condition and sex effects on cortisol concentrations (see Table 2). Because past research has suggested that the emotion, shame is positively associated with cortisol responses to social evaluative threat (S. S. Dickerson, Gruenewald, & Kemeny, 2004), the PANAS item, "ashamed" was included in the correlations. These analyses revealed no statistically significant correlations between cortisol

and ratings of PANAS shame, PANAS negative affect, PANAS positive affect, SACL stress, SACL arousal, effort, or perceptions of the audience (all |r|s < .18, all *p*s > .07). Therefore, it appears that psychological responses did not mediate differences in cortisol concentrations.

Table 2

Correlations Among Cortisol and Psychological Variables

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13
Cortisol													
1. Average	-												
PANAS Subscales													
2. Negative Affect	.127	-											
3. Positive Affect	-.149	-.338**	-										
SACL Subscales													
4. Stress	.04	.780**	-.432**	-									
5. Arousal	-.034	-.007	.400**	-.055	-								
6. Effort Engagement Index	-.136	-.252**	.458**	-.203**	.328**	-							
Perceptions of audience													
7. Attentive	-.043	-.046	0.094	-.167*	.014	.051	-						
8. Cheerful	.025	-.147*	.171*	-.161*	.015	-.008	.420**	-					
9. Supportive	.047	-.163*	.146*	-.216**	-.006	-.003	.341**	.752**	-				
10. Stressful	.046	.448**	-.161*	.451**	.035	-.059	-.083	-.119	-.156*	-			
11. Could Hear Me	.027	.059	.133	-.066	-.004	.116	.440**	.09	.108	.094	-		
12. Judgmental	-.046	.322**	-.220**	.371**	-.085	-.061	-.238**	-.288**	-.331**	.334**	-.094	-	
12. Sleepy	.127	.108	-.023	.215**	.071	.018	-.456**	-.148*	-.105	.034	-.323**	.299**	
13. **PANAS Item: Ashamed**	.106	.774**	-.338**	.553**	.033	-.292**	.021	-.063	-.069	.341**	.081	.267**	.086

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

3.3. Cardiovascular Measures

First, paired samples t-tests revealed that blood pressure levels during the speech (M_{SBP} = 156.10mmHg, SD_{SBP} = 24.01mmHg; M_{DBP} = 93.94mmHg, SD_{DBP} = 15.34mmHg) and math (M_{SBP} = 153.97mmHg, SD_{SBP} = 23.69mmHg; M_{DBP} = 93.23mmHg, SD_{DBP} = 14.78mmHg) tasks did not differ significantly (SBP, t = 1.47; DBP, t = .82, *p*s > .10). Therefore, speech and math task blood pressure values were averaged into a combined task value for the following analyses (Table 1). A paired t-test comparing HR during the speech and math tasks indicated that HR during the speech task was significantly higher than during the math task (Table 1), t = 3.49, $p < .01$. Therefore, HR during the speech and math tasks were analyzed independently.

3.3.1. Systolic Blood Pressure

The primary mixed-model ANOVA for SBP indicated that saliva contribution explained a significant amount of variance $F(4,384) = 6.21$, $p < .001$, $\eta^2 = .06$. Also, the analysis indicated that there was a between-subjects effect of sex $F(1,96) = 9.45$, $p < .004$, $\eta^2 = .09$ such that males had higher average SBP (M = 144.45 mmHg; SD = 16.25) than females (M = 134.4 mmHg; SD = 16.59). Follow-up pairwise LSD comparisons indicated that males had higher SBP than females at all timepoints (Table 3; *p's* < .02). Furthermore, a significant within-subjects effect indicated significant changes in SBP over the course of the study session (Table 1), $F(4,384) = 23.18$, $p < .001$, $\eta^2 = .19$. Pairwise LSD comparisons indicated that SBP differed across all study time points (*p*'s < .05) with one exception. Average SBP during the speech preparation period and Recovery 1 were not statistically different (Table 1). There was no indication of a statistically significant interaction between SBP pattern and condition, $F(8,384) = 1.63$, $p > .05$, $\eta^2 = .03$, or sex $F(4,384) = .90$, $p > .05$, $\eta^2 = .009$ (Figure 3). Additionally, there was no indication of a three-way interaction between SBP patterns, condition, and sex, $F(8,384) = 1$, $p > .05$, $\eta^2 = .02$.

Table 3

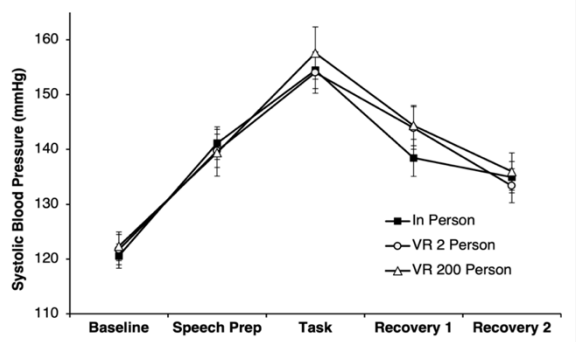
Systolic and Diastolic Blood Pressure by Study Period and Sex

Blood Pressure	Sex	Baseline	Speech Prep	Task	Recovery 1	Recovery 2
SBP (mmHg)	Male	125.31 (14.23)	145.75 (18.74)	161.94 (20.78)	148.55 (21.64)	140.4 (17.45)
	Female	118.53 (14.11)	136.07 (19.80)	150.26 (23.02)	136.8 (19.60)	130.6 (17.46)
DBP (mmHg)	Male	76.62 (9.50)	86.1 (10.84)	97.45 (12.45)	89.06 (12.74)	85.45 (11.61)
	Female	70.8 (8.00)	80.8 (12.00)	91.3 (15.14)	83.62 (11.95)	79.6 (11.01)

Note. Mean systolic and diastolic blood pressure by study period and sex. The values in parentheses represent standard deviation (SD) values. SBP = Systolic Blood Pressure; DBP = Diastolic Blood Pressure.

Figure 3

Mean Systolic Blood Pressure for TSST Conditions



Note. Error bars represent the standard error of the mean.

3.3.2. Diastolic Blood Pressure

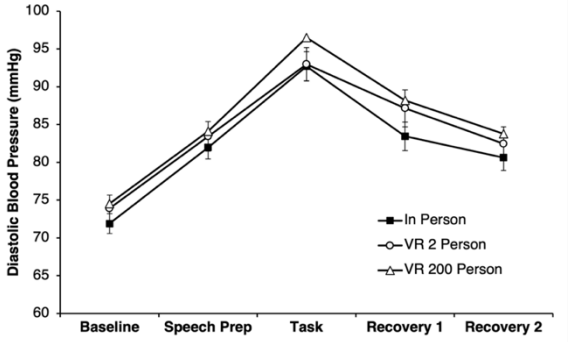
The primary mixed model ANOVA for DBP indicated that there was a between-subjects effect of sex, $F(1,97) = 7.94$, $p < .007$, $\eta^2 = .08$, such that males (M = 87.05 mmHg, SD = 10.41) had higher DBP on average than females (M = 81.14 mmHg, SD = 10.66). Follow-up pairwise LSD comparisons indicated that males had higher DBP than females at all timepoints (Table 3; *p's* < .03). Also, there were statistically significant changes in DBP over the course of the study session (Table 1), $F(4,388) = 144.24$, $p < .001$, $\eta^2 = .6$ (Table 1). Pairwise LSD comparisons revealed that across all study timepoints DBP significantly differed (all *p*s < .05) with one exception. Average DBP during the preparation period and Recovery 2 period (Table 1) did not significantly differ. There was no indication of a statistically significant interaction between the pattern of DBP and condition, $F(8,388) = .640$, $p > .05$, $\eta^2 = .01$ (Figure 4). There were no other statistically significant effects, *p*s > .05.

3.3.3. Heart Rate

The primary mixed model ANOVA on HR showed statistically significant changes in HR over the course of the study session, $F(5,485) = 85.41$, $p = .001$, $\eta^2 = .47$ (Table 1). Pairwise comparisons using the LSD method revealed that across all study timepoints HR significantly

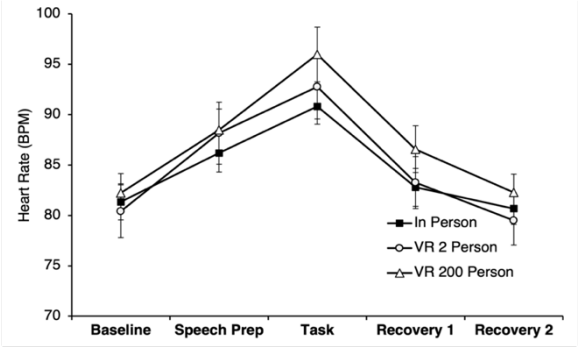
differed (all *p*s < .05) with one exception. Average HR during the baseline period and Recovery 2 period (Table 1) did not significantly differ ($p > .05$). No statistically significant condition effects on the pattern of HR pattern were found, $F(10,485) = 1.11$, $p > .30$, $\eta^2 = .02$ (Figure 5). There were no other statistically significant effects, *p*s > .05.

Figure 4
Mean Diastolic Blood Pressure for TSST Conditions



Note. Error bars represent the standard error of the mean.

Figure 5
Mean Heart Rate for TSST Conditions



Note. Error bars represent the standard error of the mean.

3.4. Psychological Measures

Means and standard deviations of the SACL, PANAS, and Effort and Engagement Index variables are shown in Table 4.

Table 4

PANAS and SACL Totals by Condition

Audience	Negative Affect*	Positive Affect*	SACL Stress	SACL Arousal	Effort & Engagement
In Person	**27.79$_{a}$ (9.33)**	**22.50$_{a}$ (5.73)**	**29.37$_{a}$ (6.31)**	**23.09$_{ab}$ (3.50)**	**5.01$_{a}$ (1.50)**
Male	26.00 (10.30)	24.59 (5.58)	25.89 (6.12)	23.41 (3.41)	4.54 (1.63)
Female	29 .00 (8.52)	21.08 (5.46)	31.73 (5.33)	22.88 (3.59)	5.32 (1.29)
VR 200	**28.21$_{a}$ (8.49)**	**22.86$_{a}$ (8.14)**	**30.05$_{a}$ (6.73)**	**23.93$_{b}$ (2.13)**	**5.07$_{a}$ (1.30)**
Male	28.09 (9.22)	25.39 (8.63)	28.35 (7.16)	24.65 (3.29)	5.11 (1.67)
Female	28.29 (8.12)	22.86 (8.14)	31.21 (6.27)	23.44 (2.39)	5.05 (1.02)

Note. PANAS and SACL totals by condition and sex. The values in parentheses represent standard deviation (SD) values.
[a]SACL = Stress and Arousal Checklist. * = PANAS subscale.
[b]Different subscripts within columns denote statistically significant group mean differences, *p*s < .05.

3.4.1. Stress and Arousal Checklist

A 3 (audience condition) x 2 (participant sex) ANOVA revealed no statistically significant condition differences in the SACL Stress subscale, $F(2,197) = 1.87$, or the SACL Arousal subscale, $F(2,197) = 1.62$, both *p*s > .10. A significant sex effect was found for both the SACL Stress subscale, $F(1,197) = 16.21$, $p < .001$, $\eta^2 = .078$, and the SACL Arousal subscale, $F(1,197) = 4.39$, $p < .04$, $\eta^2 = .02$. Females had higher stress scores (Male M = 26.7, SD = 6.59; Female M = 30.51, SD = 6.52) and males had higher arousal scores (Male M = 23.88, SD = 3.41; Female M = 22.81, SD = 3.44).

3.4.2. Positive and Negative Affect Schedule

Independent analyses of the PANAS positive and negative affect scales showed that there was a significant condition effect on negative affect, $F(2,197) = 4.29$, $p < .02$, $\eta^2 = .04$. Post hoc LSD analyses indicated that the VR 2 condition elicited lower general negative affect scores than the IP 2 or VR 200 conditions (see Table 4). There was no statistically significant condition effect on the general positive affect scale of the PANAS, $F(2,197) = .13$, $p > .90$.

However, there was a significant effect of sex on positive affect, $F(1,197) = 11.21$, $p < .002$, $\eta^2 = .06$, such that males reported higher positive affect than females (Male M = 24.84, SD = 7.7; Female M = 21.53, SD = 6.24). There were no other significant effects (ps > .05).

3.4.3. Effort and Engagement

The 3 x 2 ANOVA revealed no statistically significant differences in the effort and engagement index values across conditions or sexes (ps > .05, Table 4).

3.4.4. Perceptions of Audience

Means, standard deviations, and condition effects on participant's perceptions of their audiences are shown in Table 5. ANOVA analyses indicated that the VR 200 audience was perceived to be less "attentive" and more "sleepy" than the other two conditions. The VR 200 audience was also rated to be more "supportive" than the IP 2 audience. The VR 2 audience was seen as less "stressful" and sleepier than the IP 2 audience. Furthermore, males rated their audiences as more "attentive" and more "cheerful" than females. Participants rated both VR conditions as less able to hear the participant than the IP audience. There were no other significant effects (ps > .05, see Table 5).

3.5. Sex Differences in VR vs. IP

To directly address the hypothesis that sex differences in response to an IP audience would differ from sex differences in response to a pre-recorded VR audience, the VR 2 and VR 200 conditions were collapsed. Then a mixed-model 2 (IP vs VR) x 2 (male vs female) x 3 (cortisol timepoint) ANOVA analysis, with timepoint as the repeated measure, was performed. The analysis of cortisol revealed a significant 3-way interaction, $F(2,272) = 4.06$, $p < .02$, $\eta^2 = .03$. Pairwise comparisons indicated that males had higher Recovery 2 cortisol concentrations in the IP 2 condition compared to the collapsed VR conditions ($p < .05$, see Table 6).

Table 5

Audience Perception by Condition

Audience	Attentive	Cheerful	Supportive	Stressful	Judgmental	Sleepy	Could Hear Me
In Person	**3.01$_{a}$ (1.29)**	**1.40$_{a}$ (0.79)**	**1.36$_{a}$ (0.77)**	**3.23$_{a}$ (1.42)**	**3.03$_{a}$ (1.43)**	**1.52$_{a}$ (0.88)**	**4.15$_{a}$ (0.96)**
Males	3.26 (1.43)	1.67 (1.04)	1.52 (1.00)	2.76 (1.36)	2.85 (1.43)	1.59 (0.97)	4.04 (1.02)
Females	2.85 (1.17)	1.23 (0.53)	1.25 (0.60)	3.52 (1.40)	3.15 (1.42)	1.48 (0.82)	4.23 (0.92)
VR 2	**2.64$_{a}$ (1.12)**	**1.45$_{a}$ (0.89)**	**1.58$_{ab}$ (0.88)**	**2.66$_{b}$ (1.27)**	**3.07$_{a}$ (1.45)**	**2.23$_{b}$ (1.35)**	**3.23$_{b}$ (1.13)**
Males	3.04 (1.22)	1.46 (0.71)	1.52 (0.80)	2.52 (1.22)	2.93 (1.24)	2.41 (1.45)	3.52 (1.22)
Females	2.41 (1.00)	1.44 (1.00)	1.61 (1.00)	2.74 (1.31)	3.15 (1.56)	2.13 (1.29)	3.22 (1.07)
VR 200	**2.25$_{b}$ (0.97)**	**1.60$_{a}$ (0.92)**	**1.74$_{b}$ (0.97)**	**2.91$_{ab}$ (1.27)**	**3.05$_{a}$ (1.22)**	**3.35$_{c}$ (1.28)**	**3.09$_{b}$ (1.06)**
Males	2.57 (1.08)	1.96 (1.19)	2.00 (1.15)	3.00 (1.27)	2.70 (1.19)	3.00 (1.35)	3.35 (1.03)
Females	2.03 (0.83)	1.35 (0.60)	1.60 (0.90)	2.85 (1.28)	3.29 (1.19)	3.59 (1.18)	2.91 (1.06)

Note. Audience perception by condition and sex. The values in parentheses represent standard deviation (SD) values.

*Different subscripts within columns denote statistically significant group mean differences, ps < .05.

There were no condition differences in Baseline, Recovery 1, and Recovery 2 cortisol concentrations for males or females (*p*s > .05, see Table 6). Additionally, pairwise comparisons showed that males had statistically different cortisol concentrations at each timepoint in both IP and VR conditions (*p*s < .05) indicating significant reactivity and recovery. In contrast, female baseline cortisol concentrations did not statistically differ from Recovery 1 concentrations in the IP 2 condition ($p = .68$), suggesting that females did not experience significant reactivity in this condition. However, in the collapsed VR conditions, there was some indication of female cortisol reactivity as female baseline cortisol was lower than the Recovery 1 cortisol concentrations, however this difference was only marginal ($p = .054$). Also, for females, in both the IP 2 and VR collapsed conditions, Recovery 2 values were significantly lower than

Recovery 1 values ($p < .05$). Parallel analyses run on blood pressure and heart rate data revealed no significant 3-way interaction effects, (ps > .05).

Sex differences in psychological responses to the IP condition compared with the combined VR conditions were examined with 2 (VR vs IP) x 2 (male vs female) ANOVAs. The only significant condition by sex interaction was found for the effort and engagement scale $F(1,197) = 3.76$, $p = .05$, $\eta^2 = .02$. Paired LSD comparisons revealed that in the IP 2 condition, females reported higher effort and engagement than males ($p < .05$). However, in the VR condition, there was no sex difference in reported effort or engagement ($p > .05$, see Table 6).

Table 6

Cortisol and Effort and Engagement by Sex and Audience Presentation

Condition	Sex	Baseline Cortisol	Recovery 1 Cortisol	Recovery 2 Cortisol	Effort & Engagement
In-Person	Male	0.20 (0.11)	0.49 (.23)	0.41 (.20)	4.54 (1.63)
	Female	0.22 (0.12)	0.23 (.12)	0.18 (.07)	5.32 (1.30)
VR	Male	0.21 (0.12)	0.42 (.26)	0.30 (.18)	5.06 (1.52)
	Female	0.20 (0.13)	0.24 (.19)	0.21 (0.16)	5.02(1.21)

Note. Cortisol and Effort and Engagement by condition and sex. The values in parentheses represent standard deviation (SD) values.

[a]VR = Collapsed VR 200 and VR 2 Conditions.

3.6. Post hoc: Cortisol Sex Differences in VR 2 vs VR 200

Although there were no a priori hypotheses regarding sex differences in responses to the VR 2 compared to VR 200 conditions, results of prior analyses indicated that average cortisol concentrations in the VR 200 conditions were significantly greater than those in the VR 2 condition (see above). Therefore, sex differences were explored by comparing cortisol reactivity and recovery in the VR 2 and VR 200 conditions with separate 2 (VR 2 vs VR 200) x 3 (cortisol timepoint) mixed model ANOVAs for each sex. Once again, cortisol timepoint was the repeated measure, and VR condition was the between-subjects factor. Results of these

analyses confirmed that males had significant changes in cortisol concentrations, $F(2,68) = 22.13$, $p < .001$, $\eta^2 = .39$, with significant reactivity and recovery (LSD ps < .05). There were no other significant effects for males (ps > .05). Females had significant changes in cortisol concentrations, $F(2,110) = 3.33$, $p < .05$, $\eta^2 = .06$ with a significant increase from Baseline (M = 0.18 SD = .10) to Recovery 1 (M = .21, SD = .14), $p < .05$, a significant decrease from Recovery 1 (M = .21, SD = .14) to Recovery 2 (M = 0.18, SD = .12), $p < .001$, and no difference between Baseline (M = 0.18 SD = .10) and Recovery 2 (M = 0.18, SD = .12), $p > .05$. In addition, for females there was a marginally significant between-subjects effect of condition on cortisol concentrations ($p = .07$) with greater cortisol concentrations in the VR 200 (M = .22, SD = .21) compared to the VR 2 condition (M = .17, SD = .17).

4. DISCUSSION

On average, the TSST elicited significant cortisol and cardiovascular increases (i.e., reactivity) followed by significant decreases (i.e., recovery). There was some indication of moderating effects of condition and participant sex on physiological and psychological responses. In the following discussion each of the hypotheses are considered separately.

4.1. Hypothesis 1: Sex Differences in Physiological Responses

There was mixed support for the first hypothesis, that males would have greater physiological reactivity to the TSST. In general, males and females had uniform and significant cortisol reactivity (marginally significant for females) and recovery. However, while male and female baseline cortisol did not statistically differ, males had higher cortisol concentrations than females at Recovery 1 and Recovery 2 in all conditions. This is consistent with previous findings that males had higher cortisol than females at peak (Recovery 1) and recovery (Recovery 2) periods, but not at baseline (J. Liu et al., 2017). Thus, it appears that sex differences in cortisol concentrations in response to the TSST are a result of reactivity as opposed to basal HPA-axis functioning.

In terms of cardiovascular measures, average levels of blood pressure were higher in males. In general, males had significantly higher levels of SBP and DBP at every timepoint ($ps < .05$). This is consistent with findings that males tend to have higher blood pressure than premenopausal females (J. F. Reckelhoff, 2001). The patterns of SBP and DBP change over the course of the study did not differ by sex, suggesting that the magnitudes of reactivity and recovery were similar. Heart rate did not significantly differ by sex at any timepoint, suggesting that male and female differences in blood pressure were due to differences in blood pressure components other than heart rate (e.g., stroke volume, peripheral resistance; Girdler,

Turner, Sherwood, & Light, 1990). Future research might consider how these individual components of blood pressure may contribute to sex differences in cardiovascular disease rates across the lifespan.

4.2. Hypothesis 2: Sex Differences in Psychological Responses

The results indicated mixed support for the second hypothesis, that females would report greater psychological stress responses than males. In general sex differences in psychological responses to the TSST fell along two dimensions with females reporting more stress and males reporting more positive affect and arousal. Consistent with feeling more positive affect, males also reported perceiving audiences as more supportive and more cheerful.

It is possible that females reported more negative responses than males due to sex differences in the socialization of self-efficacy and self-image (Bacchini & Magliulo, 2003), possibly influencing psychological perspectives in perceptions of the TSST (i.e., males more confident and females more cautious) (Behnke & Sawyer, 2000). Specifically, males may perceive the TSST to be a challenge and thus result in more positive affect, arousal, and thinking the audience was more supportive and more cheerful. Conversely females may have perceived the TSST to be threatening and therefore more stressful, less positive, and the audience less supportive and less cheerful. Future research may want to ask participants how they perceived the speech and math task components of the TSST specifically, rather than the psychological affects alone, to explore how males and females perceived the stress tasks.

4.3. Hypothesis 3: Audience Size Effects

The results provided support for the third hypothesis, that a large VR 200-person audience would elicit greater psychophysiological responses than a smaller VR 2-person audience. In examining the effect of a large, VR 200-person audience, the VR 200 condition

elicited average cortisol concentrations and ratings of negative affect that were greater than those in the VR 2 condition.

When the VR 200 condition was compared to the IP 2 condition, results revealed that these conditions elicited similar psychophysiological responses. However, the audiences in these conditions were perceived differently. The VR 200 audience was seen as more sleepy and less attentive than the IP 2 audience. This is likely due, at least in part, to a combination of factors. First, whereas the IP 2 audience's negative responses could be coordinated with participant's behaviors including specific statements or calculation errors, the pre-recorded audience behaviors were consistent across participants and did not depend on the participant's performance. Such, non-contingent behaviors may be perceived as reflecting less attention and more sleepiness.

However, because the VR 2 audience was not perceived to be significantly sleepier or more inattentive than the IP 2 audience, it is more likely that the VR 200 vs IP 2 differences in perceived sleepiness and inattentiveness were due to the increased size of the audience. That is, the larger audience size may have increased the perception of sleepiness and inattention due to the larger number of individuals displaying relevant behaviors (e.g., looking around the room). Future research should consider the relative contributions of non-contingent audience behaviors and audience size to perceptions of audience sleepiness and attentiveness.

Surprisingly, the sleepier, less attentive VR 200 audience was also perceived to be more supportive than the IP 2 audience. It is possible that non-contingent sleepy and inattentive behaviors of the VR 200 audience are less *non*-supportive than the contingent, attentive behaviors of the IP 2 audience. Consistent with this, the VR 2 condition received an intermediate, though not significantly different rating of audience supportiveness. The VR 2

audience's non-contingent behaviors may have been perceived as more supportive (or less critical) than the IP 2-person audience's contingent behaviors. Furthermore, the perception of the audience's non-positive behavior may have been diffused by the VR 200 condition's larger audience size (Alt & Phillips, 2021; Phillips, Weisbuch, & Ambady, 2014). The participants may have believed that some of the 200 audience members may have been evaluating them positively (i.e., supportively) even if those audience members were not specifically identifiable. Future research should explore what audience qualities or characteristics individuals attend to when audience size is variable.

4.4. Hypothesis 4: IP vs. VR Sex Differences

The results indicated mixed support for the fourth hypothesis, that sex differences in response to the VR-TSST would be different from sex differences in response to the IP-TSST. Although the primary analysis did not reveal a condition by sex by cortisol timepoint three-way interaction, sex differences were explored with focused, hypothesis-driven analyses comparing male and female responses to the IP 2 vs a collapsed VR condition, and by examining male and female responses to conditions separately. In the analyses males showed consistent cortisol reactivity and recovery in all conditions. For females, while primary analyses suggested marginally significant increases in cortisol (i.e., reactivity), when analyses were restricted to only the VR conditions (Post Hoc VR 2 vs VR 200 analyses) there was a significant increase in cortisol. This suggests that the full model analyses may have lacked power (observed power for a 3-way interaction = .644), warranting caution in our interpretation of sex differences in this analysis.

Nevertheless, the analyses suggest that while males had uniformly statistically significant cortisol reactivity across all conditions, females had statistically significant reactivity

to the VR conditions (in focused analyses) but not in the IP 2 condition. Furthermore, there was some indication that females produced greater cortisol concentrations, on average, in the VR 200 condition compared to the VR 2 condition. In terms of cortisol concentrations, females may be more sensitive than males to the TSST in VR and the increase in audience size. Furthermore, it suggests that male-female cortisol response differences in IP conditions may be slightly exacerbated. That is, perhaps audiences in IP conditions respond differently to males and females, increasing the likelihood that males will show greater cortisol reactivity than females. In contrast, under more controlled audience conditions in VR this sex difference in cortisol reactivity may be somewhat mitigated (i.e., both males and females showed significant reactivity).

Differences in psychological responses to IP and VR conditions by males and females further support this idea. Specifically, males and females did not differ in their reports of effort and engagement in a collapsed VR condition, but females reported putting forth more effort and being more engaged than males in the IP condition. It is possible that this difference in effort and engagement was the result of differences in audience behaviors. For instance, audiences tend to perceive male performers as more competent than female performers (Aalberg & Jenssen, 2007). If this perception led to noticeable differences in audience behavior, female participants may have felt they had to work harder to elicit the IP audience's approval.

In terms of cardiovascular responses, both males and females had similar patterns of blood pressure and heart rate responses to IP and VR conditions (i.e., increases and decreases). However, males had higher blood pressure than females at all timepoints while heart rate did not differ by sex. This is consistent with cardiovascular research, such that males have higher

blood pressure and greater risk of cardiovascular disease than females until menopause or middle age (Maas & Appelman, 2010; Jane F Reckelhoff, 2001).

4.5. Limitations

One of the aims was to compare the use of a VR-TSST protocol to a "control" condition (i.e., a replication of the traditional in-person TSST). Other comparison conditions may have revealed additional effects of using a VR protocol. For instance, this study was unable to assess the effect of simply wearing an Oculus headset on physiological functioning. It is possible that wearing a headset may have increased anxiety or stress in a way that an IP audience did not. Also, the VR protocol was designed to mimic an IP TSST protocol as closely as possible, including leading participants to believe that the pre-recorded audience was live. It's not clear if this is essential for the effects found here and should be examined by future studies.

Also, in the present study, the pre-recorded VR 200-person audience was not compared to an IP 200-person audience. Repeatedly employing an IP audience of 200 persons for each participant was not feasible due to space, time, and continuity constraints. Thus, it is not clear if an IP audience of this size would elicit a greater stress response than the pre-recorded audience. Questions involving these and other parameters introduced by using a VR-TSST should be investigated by future research. There are limitations to using a pre-recorded audience stimulus. The resolution of the 3D, 360° recording did not capture details for faces in the back rows of the 200-person audience. It is possible that this lack of acuity lessened the impact of the large audience (Mostajeran et al., 2020). Also, using a pre-recording of the audience risks anachronistic cues such as an audience in summer attire viewed by a participant in winter months. Such incongruities may mitigate the impact of the audience. Future studies that employ a pre-recorded audience should consider these variables. In the examination of sex differences

in response to the VR-TSST, there may have been sex differences in a participant's familiarity with, and past use of VR technology (e.g., head mounted displays). VR technology has been utilized in videogame entertainment, generally considered to be a male-dominated space, meaning that males may have been more comfortable and familiar with VR than females thus eliciting less stress in males (Heron, Belford, & Goker, 2014). It is, however, unclear if sex differences in exposure to and use of VR contributed to psychological response differences between males and females. Future studies utilizing a VR-TSST should consider participant exposure to VR (e.g., if it is a novel experience).

Data collection for this study was halted prior to the Covid-19 pandemic, limiting the number of participants in each condition. As this was the case, there were power limitations in the comparison of sex differences and conditions. Specifically, there were only 16 male participants in IP 2 condition. It is however notable that although there were fewer males than females in this sample, results concerning male physiological responses were consistently robust and significant.

4.6. Conclusion

Classic iterations of the TSST are limited by laboratory resources and individual confounds of human consistency. This study suggests that a pre-recorded immersive audience can be used to conduct a virtual version of the TSST that elicits physiological stress reactivity like that elicited by an IP audience in both male and female participants. In general, males and females had similar physiological reactivity to the TSST. However, when considering sex and cortisol concentrations alone, males but not females had robust and significant changes in cortisol concentrations in response to all versions of the TSST. It is possible that changes in cortisol may be a better indicator of stress responsivity in males than in females.

In this study, results indicated that a larger VR 200 audience elicited greater physiological reactivity than a smaller VR 2 audience. Also, the VR 200 condition elicited cortisol concentrations more like the IP 2 condition compared to the VR 2 condition. This might suggest that overall, larger audiences elicit greater stress reactivity compared to smaller audiences in VR. Compared to the IP-TSST, males and females had more similar physiological and psychological responses to the VR-TSST. The enhanced consistency of the prerecorded audience may have eliminated confounds of participant sex specific audience behaviors, resulting in similar psychophysiological responses to the TSST. More research is needed focusing on how people (i.e., an audience) respond to male and female speakers to further inform these sex effects. Overall, the TSST elicits psychophysiological stress responses in its participants in-person and more recently through virtual reality, where male and female responses are more similar. A VR-TSST therefore allows psychophysiology researchers to examine stress responsivity to previously untenable environments in males and females more equitably.

REFERENCES

Aakster, C. (1974). Psycho-social stress and health disturbances. *Social Science & Medicine (1967), 8*(2), 77-90.

Aalberg, T., & Jenssen, A. T. (2007). Gender Stereotyping of Political Candidates. *Nordicom Review, 28*(1), 17-32. doi:10.1515/nor-2017-0198

Alt, N. P., & Phillips, L. T. (2021). Person Perception, Meet People Perception: Exploring the Social Vision of Groups. *Perspect Psychol Sci*, 17456916211017858. doi:10.1177/17456916211017858

Bacchini, D., & Magliulo, F. (2003). Self-Image and Perceived Self-Efficacy During Adolescence. *Journal of Youth and Adolescence, 32*(5), 337-349. doi:10.1023/a:1024969914672

Behnke, R. R., & Sawyer, C. R. (2000). Anticipatory anxiety patterns for male and female public speakers. *Communication Education, 49*(2), 187-195. doi:10.1080/03634520009379205

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27-41. doi:10.1016/j.jpubeco.2016.11.006

Bosch, J. A., De Geus, E. J. C., Carroll, D., Goedhart, A. D., Anane, L. A., Veldhuizen Van Zanten, J. J., . . . Edwards, K. M. (2009). A General Enhancement of Autonomic and Cortisol Responses During Social Evaluative Threat. *Psychosomatic medicine, 71*(8), 877-885. doi:10.1097/psy.0b013e3181baef05

Burford, N., Webster, N., & Cruz-Topete, D. (2017). Hypothalamic-Pituitary-Adrenal Axis Modulation of Glucocorticoids in the Cardiovascular System. *International Journal of Molecular Sciences, 18*(10), 2150. doi:10.3390/ijms18102150

Dickerson, S. S. (2008). Emotional and physiological responses to social-evaluative threat. *Social and Personality Psychology Compass, 2*(3), 1362-1378. doi:10.1111/j.1751-9004.2008.00095.x

Dickerson, S. S., Gruenewald, T. L., & Kemeny, M. E. (2004). When the social self is threatened: Shame, physiology, and health. *Journal of Personality, 72*(6), 1191-1216.

Dickerson, S. S., & Kemeny, M. E. (2004). Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research. *Psychological Bulletin, 130*(3), 355-391.

Dickerson, S. S., Mycek, P. J., & Zaldivar, F. (2008). Negative social evaluation, but not mere social presence, elicits cortisol responses to a laboratory stressor task. *Health Psychol, 27*(1), 116-121. doi:2008-00647-016 [pii] 10.1037/0278-6133.27.1.116

Fallon, M. A., Careaga, J. S., Sbarra, D. A., & O'Connor, M.-F. (2016). Utility of a virtual Trier Social Stress Test: Initial findings and benchmarking comparisons. *Psychosomatic Medicine, 78*(7), 835-840.

Garfin, D. R., Thompson, R. R., & Holman, E. A. (2018). Acute stress and subsequent health outcomes: A systematic review. *Journal of Psychosomatic Research, 112*, 107-113. doi:https://doi.org/10.1016/j.jpsychores.2018.05.017

Gerardi, M., Cukor, J., Difede, J., Rizzo, A., & Rothbaum, B. O. (2010). Virtual reality exposure therapy for post-traumatic stress disorder and other anxiety disorders. *Current psychiatry reports, 12*(4), 298-305.

Gervasio, J., Zheng, S., Skrotzki, C., & Pachete, A. (2021). The effect of oral contraceptive use on cortisol reactivity to the Trier Social Stress Test: A meta-analysis. *Psychoneuroendocrinology*, 105626.

Girdler, S. S., Turner, J. R., Sherwood, A., & Light, K. C. (1990). Gender differences in blood pressure control during a variety of behavioral stressors. *Psychosom Med, 52*(5), 571-591. doi:10.1097/00006842-199009000-00009

Gonçalves, R., Pedrozo, A. L., Coutinho, E. S. F., Figueira, I., & Ventura, P. (2012). Efficacy of Virtual Reality Exposure Therapy in the Treatment of PTSD: A Systematic Review. *PLoS ONE, 7*(12), e48469. doi:10.1371/journal.pone.0048469

Goodman, W. K., Janson, J., & Wolf, J. M. (2017). Meta-analytical assessment of the effects of protocol variations on cortisol responses to the Trier Social Stress Test. *Psychoneuroendocrinology, 80*, 26-35.

Hamer, M., Endrighi, R., Venuraju, S. M., Lahiri, A., & Steptoe, A. (2012). Cortisol Responses to Mental Stress and the Progression of Coronary Artery Calcification in Healthy Men and Women. *PLoS ONE, 7*(2), e31356. doi:10.1371/journal.pone.0031356

Hellhammer, J., & Schubert, M. (2012). The physiological response to Trier Social Stress Test relates to subjective measures of stress during but not before or after the test. *Psychoneuroendocrinology, 37*(1), 119-124. doi:10.1016/j.psyneuen.2011.05.012

Helminen, E., Morton, M., Wang, Q., & Felver, J. (2019). A meta-analysis of cortisol reactivity to the Trier Social Stress Test in virtual environments. *Psychoneuroendocrinology, 110*, 104437. doi:10.1016/j.psyneuen.2019.104437

Helminen, E., Morton, M., Wang, Q., & Felver, J. (2021). Stress Reactivity to the Trier Social Stress Test in Traditional and Virtual Environments: A Meta-Analytic Comparison. *Psychosomatic medicine, 83*(3), 200-211.

Heron, M. J., Belford, P., & Goker, A. (2014). Sexism in the circuitry. *ACM SIGCAS Computers and Society, 44*(4), 18-29. doi:10.1145/2695577.2695582

Het, S., Rohleder, N., Schoofs, D., Kirschbaum, C., & Wolf, O. T. (2009). Neuroendocrine and psychometric evaluation of a placebo version of the 'Trier Social Stress Test'. *Psychoneuroendocrinology, 34*(7), 1075-1086. doi:10.1016/j.psyneuen.2009.02.008

Hilmert, C. J., Kulik, J. A., & Christenfeld, N. (2002). The varied impact of social support on cardiovascular reactivity. *Basic and Applied Social Psychology, 24*(3), 229-240. Retrieved from http://www.erlbaum.com/Journals/journals/BASP/basp.htm

Iob, E., & Steptoe, A. (2019). Cardiovascular Disease and Hair Cortisol: a Novel Biomarker of Chronic Stress. *Current Cardiology Reports, 21*(10). doi:10.1007/s11886-019-1208-7

Jansen, J., Schreuder, J., Mulier, J. P., Smith, N., Settels, J., & Wesseling, K. (2001). A comparison of cardiac output derived from the arterial pressure wave against thermodilution in cardiac surgery patients. *British journal of anaesthesia, 87*(2), 212-222.

Jönsson, P., Wallergård, M., Österberg, K., Hansen, Å. M., Johansson, G., & Karlson, B. (2010). Cardiovascular and cortisol reactivity and habituation to a virtual reality version of the Trier Social Stress Test: A pilot study. *Psychoneuroendocrinology, 35*(9), 1397-1403. doi:10.1016/j.psyneuen.2010.04.003

Kelly, M. M., Forsyth, J. P., & Karekla, M. (2006). Sex differences in response to a panicogenic challenge procedure: an experimental evaluation of panic vulnerability in a non-clinical sample. *Behaviour research and therapy, 44*(10), 1421-1430.

Kelly, M. M., Tyrka, A. R., Anderson, G. M., Price, L. H., & Carpenter, L. L. (2008). Sex differences in emotional and physiological responses to the Trier Social Stress Test. *Journal of behavior therapy and experimental psychiatry, 39*(1), 87-98.

Kelly, O., Matheson, K., Martinez, A., Merali, Z., & Anisman, H. (2007). Psychosocial stress evoked by a virtual audience: relation to neuroendocrine activity. *CyberPsychology & Behavior, 10*(5), 655-662.

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The "Trier Social Stress Test": A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology, 28*(1-2), 76-81. Retrieved from http://www.karger.ch/journals/nps/nps_jh.htm http://www.karger.com

Kivimäki, M., Virtanen, M., Elovainio, M., Kouvonen, A., Väänänen, A., & Vahtera, J. (2006). Work stress in the etiology of coronary heart disease—a meta-analysis. *Scandinavian journal of work, environment & health*, 431-442.

Knowles, E. S. (1983). Social physics and the effects of others: Tests of the effects of audience size and distance on social judgments and behavior. *Journal of personality and social psychology, 45*(6), 1263.

Kotlyar, M., Donahue, C., Thuras, P., Kushner, M. G., O'Gorman, N., Smith, E. A., & Adson, D. E. (2008). Physiological response to a speech stressor presented in a virtual reality environment. *Psychophysiology, 45*(6), 1034-1037.

Kudielka, B. M., Hellhammer, D. H., & Kirschbaum, C. (2007). Ten Years of Research with the Trier Social Stress Test--Revisited.

Kudielka, B. M., & Kirschbaum, C. (2005). Sex differences in HPA axis responses to stress: a review. *Biological psychology, 69*(1), 113-132.

Lamotte, G., Boes, C. J., Low, P. A., & Coon, E. A. (2021). The expanding role of the cold pressor test: a brief history. *Clinical Autonomic Research, 31*(2), 153-155. doi:10.1007/s10286-021-00796-4

Latané, B., & Harkins, S. (1976). Cross-modality matches suggest anticipated stage fright a multiplicative power function of audience size and status. *Perception & Psychophysics, 20*(6), 482-488.

Lemasson, A., André, V., Boudard, M., Lippi, D., & Hausberger, M. (2018). Audience size influences actors' anxiety and associated postures on stage. *Behavioural processes, 157*, 225-229.

Liu, J., Ein, N., Peck, K., Huang, V., Pruessner, J. C., & Vickers, K. (2017). Sex differences in salivary cortisol reactivity to the Trier Social Stress Test (TSST): A meta-analysis. *Psychoneuroendocrinology, 82*, 26-37.

Liu, Q., & Zhang, W. (2020). Sex Differences in Stress Reactivity to the Trier Social Stress Test in Virtual Reality . *Psychology Research and Behavior Management, Volume 13*, 859-869. doi:10.2147/prbm.s268039

Maas, A. H. E. M., & Appelman, Y. E. A. (2010). Gender differences in coronary heart disease. *Netherlands Heart Journal, 18*(12), 598-603. doi:10.1007/s12471-010-0841-y

Mackay, C., Cox, T., Burrows, G., & Lazzerini, T. (1978). An inventory for the measurement of self-reported stress and arousal. *British journal of social and clinical psychology, 17*(3), 283-284.

McEwen, B. S. (1998). Stress, adaptation, and disease. Allostasis and allostatic load. *Ann N Y Acad Sci, 840*, 33-44.

Montero-López, E., Santos-Ruiz, A., García-Ríos, M. C., Rodríguez-Blázquez, R., Pérez-García, M., & Peralta-Ramírez, M. I. (2016). A virtual reality approach to the Trier Social Stress Test: Contrasting two distinct protocols. *Behavior research methods, 48*(1), 223-232. doi:10.3758/s13428-015-0565-4

Moran, K. E., Ommerborn, M. J., Blackshear, C. T., Sims, M., & Clark, C. R. (2019). Financial Stress and Risk of Coronary Heart Disease in the Jackson Heart Study. *American Journal of Preventive Medicine, 56*(2), 224-231. doi:10.1016/j.amepre.2018.09.022

Mostajeran, F., Balci, M. B., Steinicke, F., Kühn, S., & Gallinat, J. (2020). *The effects of virtual audience size on social anxiety during public speaking.* Paper presented at the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR).

Phillips, L. T., Weisbuch, M., & Ambady, N. (2014). People perception: Social vision of groups and consequences for organizing and interacting. *Research in Organizational Behavior: An Annual Series of Analytical Essays and Critical Reviews, Vol 34, 34*, 101-127. doi:10.1016/j.riob.2014.10.001

Powers, M. B., & Emmelkamp, P. M. (2008). Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *Journal of anxiety disorders, 22*(3), 561-569.

Reckelhoff, J. F. (2001). Gender differences in the regulation of blood pressure. *Hypertension, 37*(5), 1199-1208. doi:10.1161/01.hyp.37.5.1199

Reckelhoff, J. F. (2001). Gender differences in the regulation of blood pressure. *Hypertension, 37*(5), 1199-1208.

Rothbaum, B. O., Hodges, L., Alarcon, R., Ready, D., Shahar, F., Graap, K., . . . Wills, B. (1999).

Virtual reality exposure therapy for PTSD Vietnam veterans: A case study. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies, 12*(2), 263-271.

Rowland, N. E., & Antelman, S. M. (1976). Stress-induced hyperphagia and obesity in rats: a possible model for understanding human obesity. *Science, 191*(4224), 310-312.

Ruiz, A. S., Peralta-Ramirez, M. I., Garcia-Rios, M. C., Muñoz, M. A., Navarrete-Navarrete, N., & Blazquez-Ortiz, A. (2010). Adaptation of the trier social stress test to virtual reality: psycho-physiological and neuroendocrine modulation. *J. Cyber Ther. Rehabil, 3*, 405-415.

Santl, J., Shiban, Y., Plab, A., Wüst, S., Kudielka, B. M., & Mühlberger, A. (2019). Gender Differences in Stress Responses during a Virtual Reality Trier Social Stress Test. *International Journal of Virtual Reality, 19*(2), 2-15. doi:10.20870/IJVR.2019.19.2.2912

Schutte, A. E., Huisman, H. W., van Rooyen, J. M., Oosthuizen, W., & Jerling, J. C. (2003). Sensitivity of the Finometer device in detecting acute and medium-term changes in cardiovascular function. *Blood pressure monitoring, 8*(5), 195-201.

Selye, H. (1936). A syndrome produced by diverse nocuous agents. *Nature, 138*(3479), 32-32.

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of surgery, 236*(4), 458.

Shiban, Y., Diemer, J., Brandl, S., Zack, R., Mühlberger, A., & Wüst, S. (2016). Trier Social Stress Test in vivo and in virtual reality: Dissociation of response domains. *International Journal of Psychophysiology, 110*, 47-55. doi:10.1016/j.ijpsycho.2016.10.008

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643-662. doi:10.1037/h0054651

Tan, S., & Yip, A. (2018). Hans Selye (1907–1982): Founder of the stress theory. *Singapore Medical Journal, 59*(4), 170-171. doi:10.11622/smedj.2018043

Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Tsao, C. W. (2021). Heart Disease and Stroke Statistics—2021 Update. *Circulation, 143*(8). doi:10.1161/cir.0000000000000950

Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the positive and negative affect schedule- expanded form*. Retrieved from https://www2.psychology.uiowa.edu/faculty/clark/panas-x.pdf

Westenberg, P. M., Bokhorst, C. L., Miers, A. C., Sumter, S. R., Kallen, V. L., Van Pelt, J., & Blöte, A. W. (2009). A prepared speech in front of a pre-recorded audience: Subjective, physiological, and neuroendocrine responses to the Leiden Public Speaking Task☆. *Biological psychology, 82*(2), 116-124. doi:10.1016/j.biopsycho.2009.06.005

Wiemers, U. S., Schoofs, D., & Wolf, O. T. (2013). A friendly version of the Trier Social Stress Test does not activate the HPA axis in healthy men and women. *Stress, 16*(2), 254-260.

Zibrek, K., Martin, S., & McDonnell, R. (2019). Is photorealism important for perception of expressive virtual humans in virtual reality? *ACM Transactions on Applied Perception (TAP), 16*(3), 1-19.

Zimmer, P., Buttlar, B., Halbeisen, G., Walther, E., & Domes, G. (2019). Virtually stressed? A refined virtual reality adaptation of the Trier Social Stress Test (TSST) induces robust endocrine responses. *Psychoneuroendocrinology, 101*, 186-192. doi:10.1016/j.psyneuen.2018.11.010

APPENDIX

A.1. Questionnaires

A.1.1. Positive and Negative Affect Scale ± Expanded Form

Instructions: This scale consists of a number of words and phrases that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you are feeling this way **DURING THE TASK.** Use the following scale to record your answers: (Watson & Clark, 1994)

1 = Very slightly or not at all

2 = A little

3 = Moderately

4 = Quite a bit

5 = Extremely

_____ cheerful	_____ sad	_____ active	_____ angry at self
_____ disgusted	_____ calm	_____ guilty	_____ enthusiastic
_____ attentive	_____ afraid	_____ joyful	_____ downhearted
_____ bashful	_____ tired	_____ nervous	_____ sheepish
_____ sluggish	_____ amazed	_____ lonely	_____ distressed
_____ daring	_____ shaky	_____ sleepy	_____ blameworthy
_____ surprised	_____ happy	_____ excited	_____ determined
_____ strong	_____ timid	_____ hostile	_____ frightened
_____ scornful	_____ alone	_____ proud	_____ astonished
_____ relaxed	_____ alert	_____ jittery	_____ interested
_____ irritable	_____ upset	_____ lively	_____ loathing
_____ delighted	_____ angry	_____ ashamed	_____ confident
_____ inspired	_____ bold	_____ at ease	_____ energetic
_____ fearless	_____ blue	_____ scared	_____ concentrating
_____ disgusted	_____ shy	_____ drowsy	_____ dissatisfied

A.1.2. Stress and Arousal Checklist

Instructions: Please answer each question truthfully on how you felt **during the task** by selecting the most appropriate choice. There are no right or wrong answers, and no trick questions. (Mackay et al., 1978)

Calm	Definitely No	A Little	Quite a bit	Definitely Yes
Contented	Definitely No	A Little	Quite a bit	Definitely Yes
Comfortable	Definitely No	A Little	Quite a bit	Definitely Yes
Uneasy	Definitely No	A Little	Quite a bit	Definitely Yes
Worried	Definitely No	A Little	Quite a bit	Definitely Yes
Distressed	Definitely No	A Little	Quite a bit	Definitely Yes
Uptight	Definitely No	A Little	Quite a bit	Definitely Yes
Tense	Definitely No	A Little	Quite a bit	Definitely Yes
Relaxed	Definitely No	A Little	Quite a bit	Definitely Yes
Bothered	Definitely No	A Little	Quite a bit	Definitely Yes
Active	Definitely No	A Little	Quite a bit	Definitely Yes
Vigorous	Definitely No	A Little	Quite a bit	Definitely Yes
Lively	Definitely No	A Little	Quite a bit	Definitely Yes
Sleepy	Definitely No	A Little	Quite a bit	Definitely Yes
Drowsy	Definitely No	A Little	Quite a bit	Definitely Yes
Energetic	Definitely No	A Little	Quite a bit	Definitely Yes
Alert	Definitely No	A Little	Quite a bit	Definitely Yes
Tired	Definitely No	A Little	Quite a bit	Definitely Yes
Passive	Definitely No	A Little	Quite a bit	Definitely Yes
Aroused	Definitely No	A Little	Quite a bit	Definitely Yes