Using virtual reality to study the impact of audience size on cortisol responses to the Trier Social Stress Test

Byron S. Garrett[1], Strahm M. Anna[2], Bagne G. Angela[1,3], Hilmert J. Clayton[1]

[1] Department of Psychology, North Dakota State University, Fargo, ND, USA

[2] Behavioral Sciences, Sanford Research, Sioux Falls, SD, USA

[2] Department of Obstetrics & Gynecology; Department of Pediatrics, Sanford School of Medicine,
  University of South Dakota, Sioux Falls, SD, USA

[3] Department of Psychology, Minnesota State Community and Technical College, Moorhead, MN, USA

Corresponding author: Hilmert J. Clayton, clayton.hilmert@ndsu.edu
Keywords: Trier social stress test – Virtual Reality – Cortisol – Audience Size

Abstract

Virtual Reality Trier Social Stress Test (VR-TSST) protocols have been shown to effectively elicit psychophysiological stress by having participants perform a speech and math task while viewing a virtual (non-present) audience. However, few studies have utilized VR technology to examine variables that would otherwise be difficult to manipulate in the lab. This study examined the impact of a large VR audience (i.e., 200 members) on the physiological (i.e., cortisol) and psychological responses of 140 individuals. Participants were randomly assigned to one of three conditions: an in-person (2-person) audience, VR 2-person audience, or a VR 200-person audience condition. Salivary cortisol was collected to assess physiological reactivity and recovery. Participants self-reported psychological responses to the TSST including stress, arousal, emotions, and perceptions of the audience. Results revealed that all conditions elicited stress reactivity. The VR 200-person condition resulted in greater cortisol concentrations and more negative affect than the small VR 2-person audience. Thus the effectiveness of a VR-TSST may be enhanced by the use of a larger virtual audience stimulus.

1. Introduction

Virtual reality (VR) technology allows researchers to simulate any environment and standardize its presentation. One area of psychology that can benefit from the use of VR is the study of stress and health. How individuals respond physiologically to stress has been implicated in the development of adverse health outcomes including heart disease and diabetes (McEwen, 1998; McEwen & Wingfield, 2003). Presenting a stressor with VR technology facilitates the manipulation of potentially influential situational factors, such as number of people present or audience size. Also, a VR stressor ensures consistency in the delivery of the stimulus, facilitating an equitable comparison of stress effects across groups (e.g., males and females) and studies. The present study employed VR versions of the Trier Social Stress Test (TSST) to examine the effects of having different audience sizes during the TSST. Additionally, we considered sex differences in responses to these environments. Because the VR-TSST is still a relatively new

protocol, we have also included a more traditional, in-person (IP), 2-person audience condition for comparison.

1.1. The Trier Social Stress Test (TSST)

For almost three decades laboratory research regarding psychophysiological stress responses has relied heavily on the TSST first described by Kirschbaum et al. (1993). During the TSST, each participant gives a speech and completes an oral math task in front of a confederate audience instructed to respond in a non-positive manner. This prompts acute physiological stress responses, including increases in the stress hormone cortisol (Goodman, Janson, & Wolf, 2017). Analyses have shown that subtle variations in the TSST can influence the effects of the protocol. For instance, an audience that responds negatively elicits a weaker cortisol response than a neutral audience (Goodman et al., 2017). Thus, it is important to have a consistent presentation of the TSST protocol and audience stimulus. A VR-TSST can provide this consistency.

1.2. Virtual Reality - TSST

Broadly speaking, VR refers to the computer simulation or representation of an interactable environment. Studies that have compared the psychophysiological effects of a VR-TSST to those of an IP-TSST have not been entirely consistent. Kelly et al. (2007) reported that although they elicited similar stress-appraisals, the IP audience elicited a greater cortisol response than a VR audience. In contrast, another direct comparison found that cortisol responses to a VR-TSST were comparable to an IP-TSST (Zimmer et al., 2019). Variability in VR versions of the TSST, including whether the VR was presented using a head mounted display, wall projection, or computer screen, and how photorealistic the audience and environment appeared may have played roles in these inconsistencies (Helminen, Morton, Wang, & Felver, 2019). Nevertheless, a recent meta-analysis of 16 VR to IP comparisons suggested that the VR-TSST is as effective as an IP-TSST at eliciting physiological and self-reported stress reactivity (Helminen, Morton, Wang, & Felver, 2021).

1.3. Audience Size

In general, audience size has been positively associated with participant anxiety (Latané & Harkins, 1976), social influence (Bond, 2005; Latane, 1981), and the perception of being socially judged (Knowles, 1983). Lemasson et al. (2018) found that during the performance of a play, an IP audience of 128 members elicited greater self-reported anxiety from the actors than smaller audiences of thirty members and eight members. Also, responses on a public speaking self-efficacy questionnaire indicated that participants became less confident in their public speaking abilities as audience size increased from two to one-hundred members (Hilmert, unpublished data). One study of physiological reactions to audience size demonstrated that performing a speech to an IP audience of four members elicited significantly greater heart rate (HR), pre-ejection period, and salivary cortisol compared to an in-person audience of one member (Bosch et al., 2009). Thus, larger audiences may enhance stress-related reactions.

On the other hand, Mostajeran et al., (2020) found that a smaller VR audience of three members elicited higher HR than the larger, six- and fifteen-member VR audiences. Research suggests that perceptions of groups often entail *ensemble coding,* when, for example, participants perceive the emotions of groups as an average of the emotions expressed on the individual group members' faces (Alt & Phillips, 2021; Haberman & Whitney, 2009; Phillips, Weisbuch, & Ambady, 2014). This suggests that as the TSST

audience size increases, the non-positive response of the confederate audience may be *less* intense. That is, when a small audience is present, every face is clearly visible, and a uniform non-positive disapproval is perceived. When an audience is very large, the perception of the audience's disapproval may be diffused among the 200 audience members' faces, which are not all consistently attended to.

Unfortunately, studying the physiological impact of an audience of considerable size with a traditional, IP-TSST is not feasible. However, a pre-recorded, VR audience would allow for the examination of audience size without the personnel requirements of an IP-TSST. In the present study we employed TSST protocol variations intended to be conceptual replications of the traditional TSST following current recommended guidelines (Labuschagne, Grace, Rendell, Terrett, & Heinrichs, 2019).

There were two primary predictions. First, based on the most recent meta-analysis (Helminen et al., 2021) we expected the 2-person audience VR-TSST (VR 2) condition to elicit less intense cortisol and psychological stress responses than the 2-person audience IP-TSST (IP 2). Next, relative to the VR 2 condition, we expected the 200-person audience VR-TSST (VR 200) condition to either enhance (Bosch et al., 2009) or diffuse (Alt & Phillips, 2021; Phillips et al., 2014) the impact of the VR audience on cortisol and psychological responses.

We had no hypothesis concerning the relative impacts of the IP 2 condition and the VR 200 condition and consider these analyses exploratory. That is, if the VR 200 condition elicits less intense stress responses than the VR 2 condition, then the VR 200 condition may also elicit less intense stress responses than the IP 2 condition. However, if the VR 200 condition elicits more intense stress responses, then it may be more similar to the IP 2 condition than the VR 2 condition, or it may be more extreme. Finally, because previous research has reported male and female differences in stress reactivity to IP and VR TSST protocols (J. J. W. Liu et al., 2017; Q. Liu & Zhang, 2020; Santl et al., 2019), we included sex as a factor in our analyses. Consistent with past studies, we expected males to exhibit greater cortisol responses overall, and females to report more psychological stress.

2. Methods

2.1. Participants

A total of 140 students from a midwestern university participated in this study (63% female) for course credit. All procedures were approved by the Institutional Review Board of North Dakota State University.

2.2. Procedure

Each participant scheduled a 90-minute lab appointment between 11am-5pm in order to minimize the effects of diurnal rhythm on cortisol (Dickerson & Kemeny, 2004). Participants were randomly assigned to complete a five-minute speech and five minutes of oral arithmetic in one of three audience conditions: (1) an in-person, 2-person audience condition (IP 2), (2) a VR 2-person audience condition (VR 2), or (3) a VR 200-person audience condition (VR 200).

Upon arriving, participants were seated at a small table. They were told that they were about to take part in a laboratory challenge while physiological measures were taken. After 5 minutes of instructions and walking through the consent process the participant signed an approved consent form. Then participants provided the first of three saliva samples (baseline). Participants were also fitted with physiological recording equipment (i.e., electrodes and cardiovascular biometric sensors) not relevant to the present study. Next, participants were asked to sit quietly during a 10-minute acclimation period in order to adjust to the recording equipment.

Following the baseline period, the experimenter informed the participant that they would be performing a speech for an audience. To enhance the evaluative nature of the situation participants were also told that video recordings of their performances would be analyzed by experts in public speaking. There was then a 5-minute speech preparation period during which the participant mentally prepared a 5-minute speech about why the audience should hire them for a job.

For the *IP 2 condition*, the experimenter returned and informed the participant that they were waiting for the arrival of the audience. After waiting for two minutes a 2-person audience of confederate research assistants entered the lab. They were seated in front of the participant and the experimenter asked them to observe and evaluate the speech about to be given by the participant. After being reminded that the participant was to speak for the entire 5-minute period and answering any participant questions, the experimenter started the video camera and instructed the participant to begin the speech. After the 5-minute speech participants performed a "standardized cognition task," counting backwards from 2083 by 13s as quickly and accurately as possible for 5 minutes.

In the *VR 2 and VR 200 conditions*, the experimenter informed the participant, "We are using a new technology today called a virtual reality speech conduit or VRSC for short. This technology involves using a VR headset that allows you to see your audience." Also, to enhance the evaluative nature of the VR stimuli, the experimenter added that there was, "a screen in front of the audience [that] allows them to see you." After any questions were answered the participant was fitted with an Oculus headset (Oculus VR, California) connected to a PC running Vizard 6 (WorldViz, California) software, which displayed one of two immersive, pre-recorded, 360°, three dimensional visual stimuli created for this study (see below). For the first 2 minutes of wearing the headset participants viewed an empty auditorium allowing them to orient themselves in the virtual environment. After this orienting period the experimenter informed the participant "I will now turn on the audience feed" and either a 2-person or a 200-person audience appeared in the auditorium and the participant gave their 5 minute speech and performed 5-minutes of oral arithmetic.

For *all conditions*, confederate audiences were instructed prior to the experiment session to respond to the participant's performance in an evaluative, non-positive manner (see below). At three and a half minutes into the speech the experimenter reminded the participant to speak for the entire 5-minutes. During the oral arithmetic task the experimenter pointed out incorrect calculations followed by an instruction to continue from the correct number. Experimenters maintained a non-supportive tone while reminding the participants that they needed to make calculations quickly and accurately.

After the arithmetic was completed, the experimenter excused the audience, either by asking the in-person audience members to leave or by removing the Oculus headset. Next the participant was given a questionnaire to complete during the 10-minute Recovery 1 period. A second saliva sample (Recovery 1) was taken at the end of this period. Participants then continued to sit for a 10-minute Recovery 2 period followed by the final saliva sample (Recovery 2) and debriefing.

2.3. Salivary Cortisol

To measure cortisol concentrations, saliva samples were collected using Salivettes (Germany). Three Salivettes were collected during the laboratory session. Saliva samples were collected after a 5-minute consent period, and again at 20 minutes and 35 minutes after the initiation of the speech task. According to recommendations made in the literature this provided baseline, peak, and early recovery levels of cortisol, respectively (Dickerson & Kemeny, 2004). Saliva samples were shipped to Salimetrics (CA) for cortisol assays in duplicate.

Samples were tested for salivary cortisol using a high sensitivity enzyme immunoassay (Cat. No. 1-3002). Sample test volume was 25 µl of saliva per determination. The assay has a lower limit of sensitivity of 0.007 µg/dL, a standard curve range from 0.012 - 3.0 µg/dL, and an average intra-assay coefficient of variation of 4.60%, and an average inter-assay coefficient of variation of 6.00%.

2.4. Psychosocial Measures

2.4.1. Stress and Arousal Checklist (SACL)

Following the TSST tasks participants completed the SACL (Mackay et al., 1978). Participants were asked to rate the extent to which they experienced 20 emotions (e.g., calm, lively) during the tasks using a 4-item scale ranging from "definitely no" to "definitely yes". In our sample the SACL Stress sub-scale met reliability criteria, Cronbach's α = .91 and items were averaged to create a SACL Stress Index. The SACL Arousal subscale did not show good reliability, Cronbach's α = .52. For this reason, the results involving the SACL Arousal subscale index should be interpreted with caution. Higher values on these indexes indicated greater self-reported stress and arousal during the TSST.

2.4.2. Positive and Negative Affect Scale (PANAS)

The Positive and Negative Affect Scale – Expanded form (PANAS-X) (Watson & Clark, 1994), asked participants to rate the extent to which they felt 20 emotions "during the tasks," on a scale from 1 (not at all) to 5 (extremely). For the present study we focused on the General Positive Affect and General Negative Affect Subscale Indexes. The ten items making up each index had acceptable reliability, (both α's = .87). Subscale items were averaged to create the appropriate indexes. Higher values on these indexes indicated stronger emotional responses to the TSST.

2.4.3. Effort and Engagement Index

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 with lower ratings indicating less effort and engagement. The seven items were found to be highly intercorrelated (Cronbach's α = .83). These items were combined by averaging them to create an Effort and Engagement Index in which higher values indicated the participant put forth more effort and was more engaged in the tasks.

2.4.4. 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 asked how well the audience could hear the participant. These questions were answered on a 5-point Likert scale with 1 = "not at all" and 5 = "extremely." These items were analyzed separately.

2.5. Audience Stimuli

2.5.1. In-Person Audience

For the IP 2 condition, two undergraduate confederates dressed in lab coats and carried clipboards. These audience members were trained to appear evaluative, periodically taking notes, and also 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, looking through the participant, glancing at a watch,

and slightly shaking their heads. Two-person audiences were made up of two females or a male and a female. Audiences were instructed to provide the same feedback to all participants and were blind to hypotheses.

2.5.2. Virtual Reality (VR) Audiences

The VR stimuli were immersive 360° 3D recordings of audiences created using a Vuze (Human Eyes, New York) 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 help 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 help 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. Then, as if in response, the audience waved to the camera. To mirror the IP, lab experience a confederate experimenter in a white lab coat could be seen in the VR environment standing stage right, appearing to take notes throughout the recording.

VR 2

As in the IP 2 condition, the VR 2-person 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 the VR audience appeared to be the same distance from the participant as the IP 2 audience. Equivalent to the IP audience condition, two recordings of the VR 2-person audience were made with either two female research confederates, or a male and a female confederate. Presentation of the two versions was counterbalanced. Confederates were trained to respond with disinterested and evaluative actions in the same manner as the IP 2 audience (see above).

VR 200

Students were recruited to act as the 200-person audience by offering class credit for participation. Prior to recording, instructions with a list of suggested behaviors were distributed to the audience. 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 their gestures subtle so that they did not seem disingenuous. To enhance the stressful nature of the situation a member of the audience left the theater at three minutes and eight minutes into the tasks.

2.6. Analyses Overview

First, to assess the overall impact of condition on cortisol concentrations while accounting for participant sex, 3 (condition) x 2 (sex) ANOVAs were performed on cortisol area under the curve with respect to ground (AUCg) and with respect to increase (AUCi) values (Pruessner et al., 2003). Then, a 3 (cortisol timepoint) x 3 (condition) x 2 (sex) mixed-model ANOVA in which timepoint was the repeated measure considered how condition may have affected the pattern of cortisol reactivity while accounting for sex. Because of significant skew in the cortisol data, a natural log of cortisol [ln(cortisol+1)] transform was applied prior to these analyses. Significant effects were explored with post-hoc least significant difference (LSD) analyses. Also, we tested our hypotheses with planned comparison, LSD analyses of IP 2 vs. VR 2

effects and VR 2 vs. VR 200 effects. Psychological responses were compared across conditions and sexes similarly.

3. Results

3.1. Participants

The average age of the participants was 19.33 years (SD = 2.89). The racial makeup of the sample was 85.71% "White/Caucasian," 4.29% "Black or African American," 4.29% "Asian," and 5.71% other or multi-racial. Hispanic or Latino ethnicity was reported by 3.57% of the sample.

3.2. Cortisol

3.2.1. Cortisol Timepoint Analyses

A 3 (cortisol timepoint) x 3 (condition) x 2 (sex) mixed-model ANOVA first revealed a significant effect of timepoint on cortisol concentrations, $F(2, 268) = 50.91$, $p < .001$, $\eta^2 = .28$. This was qualified by a significant interaction with sex, $F(2, 268) = 27.26$, $p < .001$, $\eta^2 = .17$, and a marginally significant three-way interaction, $F(4, 268) = 2.20$, $p < .07$, $\eta^2 = .03$. Follow-up LSD comparisons showed that, overall, there was significant cortisol reactivity, i.e., baseline concentration was significantly lower than Recovery 1 concentration (Table 1), $p < .001$. Also, Recovery 2 cortisol concentration was intermediate to and significantly different from baseline and Recovery 1 concentrations (Table 1), all p-values < .001.

When considered for males and females separately, males had statistically different cortisol concentrations at baseline (M = 0.21 µg/dL, SD = 0.11), Recovery 1 (M = 0.44 µg/dL, SD = 0.25), and Recovery 2 (M = 0.33 µg/dL, SD = 0.19), all p-values < .001. Females had a marginally significant rise in cortisol concentration from baseline (M = 0.21 µg/dL, SD = 0.13) to Recovery 1 (M = 0.25 µg/dL, SD = 0.18), $p = .06$, and a statistically significant decrease in cortisol concentration from Recovery 1 to Recovery 2 (M = 0.21 µg/dL, SD = 0.14), $p < .001$. For females, Recovery 2 concentrations did not statistically differ from baseline, $p > .50$. Comparing male and female cortisol concentrations at each timepoint, LSD analyses revealed that males and females had cortisol concentrations at baseline that did not statistically differ, $p > .50$. Males did have statistically higher cortisol concentrations at Recovery 1 and Recovery 2, all p-values < .001.

Condition (n)	Baseline	Recovery 1	Recovery 2	AUCg	AUCi
IP 2 (47)	0.21 (0.12)	0.32 (0.21)	0.27 (0.17)	12.36 (6.66)	2.82 (5.59)
VR 2 (53)	0.18 (0.12)	0.27 (0.20)	0.22 (0.16)	10.43 (6.28)	2.16 (5.93)
VR 200 (40)	0.23 (0.14)	0.38 (0.27)	0.28 (0.18)	13.80 (7.96)	3.77 (7.09)
Overall (140)	**0.21 (0.13)**	**0.32 (0.23)**	**0.25 (0.17)**	**12.04 (7.01)**	**2.84 (6.17)**

Table 1: Cortisol (µg/dl) by Study Period and condition [M (SD)]

SBP = Systolic Blood Pressure; DBP = Diastolic Blood Pressure; HR = Heart Rate

Note. Different subscripts within rows denote statistically significant mean differences, *p*s < .05.

When considered in each condition separately, LSD comparisons revealed that males had significant reactivity and recovery in all three conditions (Figure 1a), all *p*s < .05. For females, baseline cortisol concentrations did not statistically differ from Recovery 1 concentrations in the any condition (Figure 1b), *all p*-values > .10. Also, for females in all three conditions, Recovery 1 cortisol concentrations were

significantly higher than Recovery 2 concentrations (Figure 1b), all *p*-values < .05. Comparing male and female cortisol concentrations at each timepoint within each condition, the same pattern reported above was found in which baselines did not differ and males had higher Recovery 1 and Recovery 2 cortisol concentrations than females, with one exception. In the VR 200 condition, male and female Recovery 2 cortisol concentrations did not statistically differ (Figures 1a and 1b), *p* > .10.

When we considered condition effects on male and female cortisol concentrations at each timepoint, LSD analyses showed that for males there was no effect of condition on baseline or Recovery 1 values (Figure 1a), *p*s > .05. However, male Recovery 2 cortisol concentrations were higher in the IP 2 condition than the VR 2 condition, *p* < .05. For females, there were no significant condition effects on cortisol concentrations at any timepoint, all *p*-values > .05.
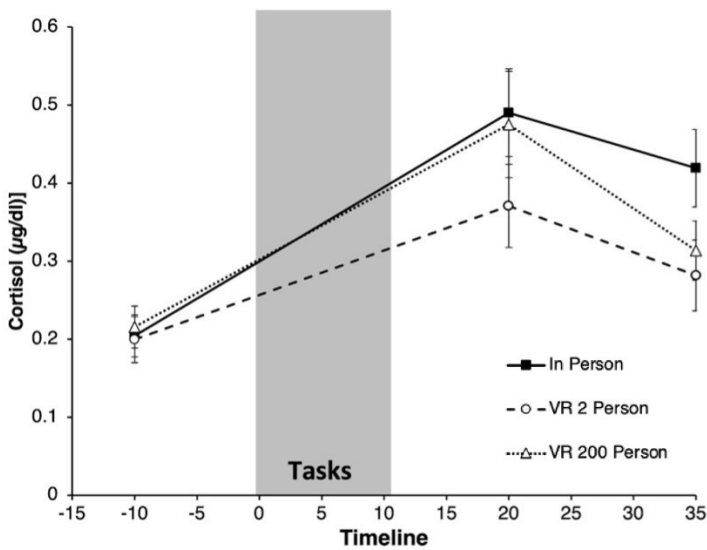

Figure 1a: Cortisol concentrations for males in responses to in-person and VR audience conditions
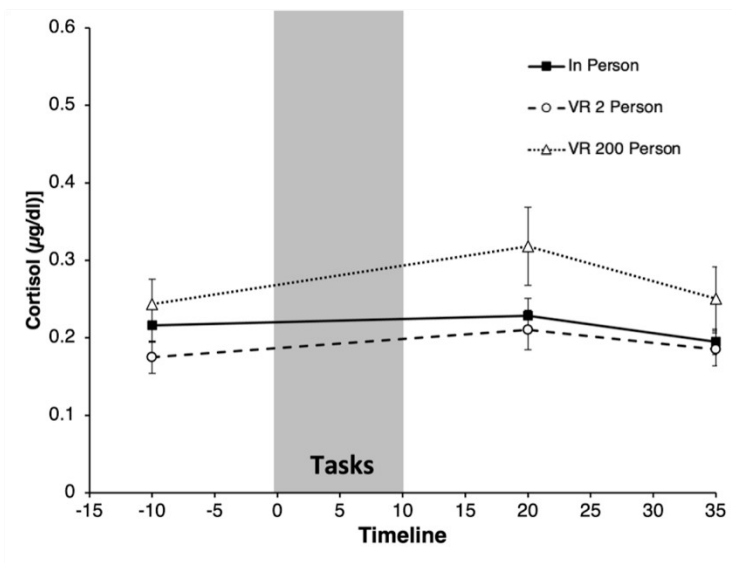

Figure 1b: Cortisol concentrations for females in responses to in-person and VR audience conditions

The analyses also revealed a marginally significant between-subjects effect of condition on average cortisol concentrations, $F_{(2,134)} = 2.90$, $p = .058$, $\eta^2 = .04$. Planned LSD comparisons showed that the VR 2 condition elicited significantly lower average cortisol concentrations (M = 0.22 µg/dL, SD = 0.13) than the VR 200 (M = 0.30 µg/dL, SD = 0.17) condition, $p < .05$. Also, the VR 2 condition elicited significantly less cortisol than the IP 2 (M = 0.27 µg/dL, SD = 0.14) condition, $p < .05$. Additionally, the analysis revealed a significant between subjects effect of sex on average cortisol concentrations, $F_{(1,134)} = 21.36$, $p < .001$, $\eta^2 = .14$ with males having higher concentrations (M = 0.33 µg/dL, SD = 0.16) than females (M = 0.22 µg/dL, SD = 0.13).

3.2.2 Cortisol AUC Analyses
Results of a 3 (condition) x 2 (sex) ANOVA on cortisol AUCg revealed a marginally significant effect of condition on cortisol AUCg, $F_{(2,134)} = 2.83$, $p = .062$, $\eta^2 = .04$. Similar to results of the prior analyses, LSD planned comparisons showed that AUCg was statistically lower in the VR 2 condition than in the VR 200 condition (Table 1), $p < .05$. Also, AUCg in the VR 2 condition was marginally statistically lower than AUCg in the IP 2 condition (Table 1), $p = .07$. Parallel analyses of cortisol AUCi, which included negative values for those that had decreases in cortisol over the course of the study (33.6% overall; 15.4% of males, 44.3% of females), did not reveal statistically significant condition effects (Table 1), all p-values $> .05$.

In terms of sex differences, analyses revealed significant effects of sex on AUCg, $F_{(1,134)} = 23.42$, $p < .001$, $\eta^2 = .15$ and AUCi, $F_{(1,134)} = 31.04$, $p < .001$, $\eta^2 = .19$. In both cases males had greater cortisol AUC (AUCg: M = 15.54, SD = 7.50; AUCi: M = 6.26, SD = 6.85) than females (AUCg: M = 9.97, SD = 5.82; AUCi: M = 0.82, SD = 4.70). There were no other significant effects on cortisol AUCg or AUCi, all p-values $> .10$.

When the use of oral contraceptives was included in analyses of female cortisol concentrations (Gervasio et al., 2021), the variable did not explain a significant amount of variance ($p > .30$) or change the results of reported analyses. There were no other significant effects involving cortisol concentrations, all p-values $> .10$.

3.3. Psychological Measures
Means and standard deviations of the PANAS, SACL, and Effort Index variables are shown in Table 2.

Condition	Negative Affect*	Positive Affect*	SACL Stress	SACL Arousal	Effort & Engagement
IP 2	27.43$_a$ (9.34)	22.15$_a$ (5.81)	29.38$_a$ (6.64)	22.98$_{ab}$ (3.53)	4.95$_a$ (1.40)
VR 2	22.96$_b$ (7.85)	24.35$_a$ (7.51)	26.96$_a$ (7.83)	22.89$_a$ (4.27)	5.14$_a$ (1.36)
VR 200	27.78$_a$ (8.95)	23.00$_a$ (7.88)	29.68$_a$ (7.33)	23.78$_b$ (2.84)	5.09$_a$ (1.36)

Table 2: Psychological Variable Totals [M (SD)] by Condition

* = PANAS subscale; SACL = Stress and Arousal Checklist

Note. Different subscripts within columns denote statistically significant overall group mean differences, ps $< .05$

3.3.1. Positive and Negative Affect Schedule

Independent 3 (condition) x 2 (Sex) ANOVAs of the PANAS positive and negative affect scales showed that there was a significant condition effect on negative affect, $F(2,135) = 4.87$, $p < .01$, $\eta^2 = .07$. Follow-up LSD analyses indicated that the VR 2 condition elicited lower general negative affect scores than the IP 2 or VR 200 conditions (see Table 2). The effect of sex and the condition by sex interaction were not statistically significant, *all p*-values > .10. For positive affect, the ANOVA revealed a significant effect of sex, $F(1,135) = 8.06$, $p < .01$, $\eta^2 = .06$, such that, on average, males reported higher positive affect (M = 25.40, SD = 7.68) than females (M = 21.96, SD = 6.48). There were no other statistically significant effects on the PANAS positive affect scale, *all p*-values > .30.

3.3.2. Stress and Arousal Checklist

Parallel ANOVA analyses of the SACL Stress subscale revealed a significant effect of sex, $F(1,135) = 9.72$, $p < .01$, $\eta^2 = .07$, with females reporting more stress on average (M = 29.96, SD = 7.14) than males (M = 26.17, SD = 7.18). The ANOVA analysis of the SACL Arousal subscale revealed a marginally significant effect of sex, $F(1,135) = 3.53$, $p < .07$, $\eta^2 = .03$, with males reporting more arousal on average (M = 23.98, SD = 3.64) than females (M = 22.71, SD = 3.59). There were no statistically significant effects of condition or of the condition by sex interaction on the SACL Stress or Arousal subscales, all *p*-values > .10.

3.3.3. Effort and Engagement

Analyses revealed no statistically significant effects of condition or sex on the effort and engagement index values, all *p*-values > .20 (Table 2).

3.3.4. Perceptions of Audience

Means and standard deviations of participants' perceptions of their audiences are shown in Table 3. Independent 3 (condition) x 2 (sex) ANOVA analyses showed that there were significant effects of condition on perceptions of the audience as "attentive," $F(2,135) = 3.51$, $p < .05$, $\eta^2 = .05$, and "sleepy," $F(2,135) = 3.87$, $p = .05$, $\eta^2 = .03$. Follow-up LSD analyses, shown in Table 3, revealed that the VR 200 audience was perceived to be less attentive and sleepier than the other two conditions. Also, the VR 2 audience was seen as sleepier than the IP 2 audience (Table 3). There was also a marginally significant effect of condition on perceptions of the audience as "stressful," $F(2,131) = 2.58$, $p = .08$, $\eta^2 = .04$. Follow-up LSD analyses revealed that the VR 2 audience was perceived to be significantly less "stressful" than the IP 2 audience. Consistent with this, planned comparisons showed that the VR 2 audience was perceived as more supportive than the IP 2 audience (Table 3). Perceptions of the VR 200 audience as stressful and supportive were intermediate to and not statistically different from perceptions of the other audiences, $p > .10$. There was a significant effect of condition on perceptions of the audience's ability to hear the participant, $F(2,135) = 10.02$, $p < .001$, $\eta^2 = .13$. Participants rated both VR conditions as less able to hear the participant than the IP 2 audience (Table 3).

There were significant effects of participant sex on perceptions of the audience as "attentive," $F(1,135) = 3.87$, $p = .05$, $\eta^2 = .03$, and "cheerful," $F(1,133) = 6.69$, $p < .05$, $\eta^2 = .05$. Males rated audiences as more attentive (M = 2.85, SD = 1.22) and more cheerful (M = 1.69, SD = 0.97) than females (Attentive: M = 2.48, SD = 1.11; Cheerful: M = 1.31, SD = 0.75). There were no other significant effects on perceptions of audiences, all *p*-values > .10.

Condition	Attentive	Cheerful	Supportive	Stressful	Judgmental	Sleepy	Could Hear Me
IP 2	2.87$_a$ (1.30)	1.23$_a$ (0.63)	1.32$_a$ (0.66)	3.31$_a$ (1.40)	2.94$_a$ (1.47)	1.51$_a$ (0.88)	4.13$_a$ (1.04)
VR 2	2.70$_a$ (1.12)	1.53$_a$ (0.95)	1.68$_b$ (0.94)	2.57$_b$ (1.28)	2.96$_a$ (1.47)	2.19$_b$ (1.35)	3.34$_b$ (1.19)
VR 200	2.22$_b$ (0.97)	1.59$_a$ (0.92)	1.61$_{ab}$ (0.97)	2.95$_{ab}$ (1.27)	3.10$_a$ (1.22)	3.34$_c$ (1.28)	2.98$_b$ (1.06)

Table 3: Audience Perception [M (SD)] by Condition

Note. Different subscripts within columns denote statistically significant group mean differences, *p*s < .05

3.3.5. Correlations with Cortisol

To explore potential mediators of the condition effects on cortisol concentrations we computed correlations between the average cortisol measures, and psychological responses to the TSST protocols (Table 4). Because past research has suggested that the emotion, shame is positively associated with cortisol responses to social evaluative threat (Dickerson et al., 2004), we included the PANAS item, "ashamed" in our correlations. 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 < .15, all *p*-values > .06). Therefore, it appears that cortisol concentrations were unrelated to psychological responses.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Cortisol														
1. Average	-													
2. AUCg	.991**	-												
PANAS Subscales														
3. Negative Affect	.108	.110	-											
4. Positive Affect	.135	.132	.311**	-										
SACL Subscales														
5. Stress	.041	.044	.798**	.448**	-									
6. Arousal	.048	.031	.040	.395**	-.040	-								
7. Effort Engagement Index	.155	.154	.255**	.456**	.211*	.317**	-							
Perceptions of audience														
8. Attentive	.005	.009	.017	.111	.102	.010	.133	-						
9. Cheerful	.067	.064	.153	.229**	.191*	.011	.041	.424**	-					
10. Supportive	.020	.019	.228**	.215*	.297**	.052	.017	.404**	.763**	-				
11. Stressful	.068	.072	.485**	.225**	.493**	.078	.063	.024	.198*	.242**	-			
12. Judgmental	.043	.041	.286**	.206*	.353**	-.050	.077	.228**	.304**	.349**	.325**	-		
13. Sleepy	.137	.155	.060	.044	.157	.101	.009	.428**	.169*	.171*	.050	.300**	-	
14. Could Hear Me	.036	.034	.094	.102	.026	.062	.114	.470**	.110	.108	.114	.088	.352**	-
15. **PANAS Item:** Ashamed	.116	.121	.780**	.262**	.552**	.118	.300**	.070	.152	.228**	.366**	.250**	.025	.099

Table 4: Correlations among cortisol and psychological variables

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

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

3.4. Male-Female vs. Female-Female Audience Dyads

In the IP 2 and VR 2 conditions audiences were made up of either a female-female (FF) or male-female (MF) dyad. There is some suggestion that the sex make-up of the TSST audience may significantly influence stress reactivity (Labuschagne et al., 2019). To see if effects of this audience dyad characteristic

persist in VR, we computed 2 (MF vs. FF) x 2 (participant sex) x 2 (IP vs. VR) ANOVAs on our outcomes.

Cortisol concentrations (AUCg, AUCi, average cortisol, cortisol at timepoint) were not significantly associated with audience dyad makeup or any interactions with this variable (all p-values > .05). A significant main effect of audience dyad makeup on negative affect ($p < .05$) was qualified by a significant interaction with IP/VR condition, $F(1,92) = 7.71$, $p < .01$, $\eta^2 = .08$. Follow-up LSD analyses revealed that the IP 2 FF audience elicited more negative affect (M = 29.54, SD = 8.47) than the IP 2 MF audience (M = 19.60, SD = 8.50), and the VR 2 FF audience (M = 22.00, SD = 8.23), both p-values < .05. There were no other significant effects of audience dyad makeup.

4. Discussion

The use of a VR-TSST protocol to elicit physiological stress responses was successful. Overall, there were significant increases in cortisol concentrations in each of the conditions of this study. Marginally significant results of omnibus analyses and planned comparisons provided some support for hypotheses. The IP 2 audience elicited greater average cortisol than the VR 2 audience, and the VR 200 audience elicited greater average and AUCg cortisol concentrations than the VR 2 audience.

Not all measures of cortisol showed condition effects. Analyses of measures that equally weighted and combined Baseline, Recovery 1, and Recovery 2 cortisol concentrations (i.e., AUCg and average cortisol) detected condition effects. In contrast, analyses of measures that considered cortisol concentrations relative to baseline (i.e., AUCi) or independently (i.e., timepoint analyses) lacked the sensitivity to detect the effects of condition. It is possible that audience size and modality of presentation variably affect reactivity and recovery of cortisol. Future research should consider possible determinants of this variability.

Consistent with our hypothesis and recent meta-analyses (Helminen et al., 2019; Helminen et al., 2021) our VR2 condition elicited significant, though, less robust stress responses relative to the equivalent IP 2 condition. The less robust effect of the VR 2 audience on average cortisol concentrations may have been due, in part, to the non-contingency of the pre-recorded VR audience's behaviors. That is, unlike the pre-recorded VR audience, the IP 2 audience was able to behave in a manner that appeared to be in response to the performance (e.g., exhibit subtle disapproving gestures after specific claims or during long pauses). This may have led to stronger negative affect, greater perceptions of the IP 2 audience as stressful, and higher cortisol concentrations. Alternatively, the non-contingent responses of the VR 2 audience may have elicited less negative affect, enhanced the perception of that audience's sleepiness, and led to relatively lower cortisol concentrations. Future research may want to focus on how audience behavior contingencies affect psychophysiological stress reactivity.

Also, consistent social psychological theory indicating that larger numbers of people are often more influential (Bond, 2005; Latane, 1981) and research comparing TSST audience sizes (Bosch et al., 2009), the large VR 200 audience elicited greater cortisol concentrations (average cortisol and AUCg) and negative affect than the small VR 2 audience. In addition, the VR 200 audience was seen as less attentive and sleepier than the VR 2 audience. Perhaps "*ensemble coding*" of the large audience diffused perceptions of attentiveness and enhanced perceptions of sleepiness (Alt & Phillips, 2021; Haberman & Whitney, 2009; Phillips et al., 2014) and having 200 people appear inattentive and sleepy during a performance is more unpleasant than a similar response from a smaller audience (Latane, 1981). Indeed, negative affect was positively correlated with perceptions of how stressful the audience was ($r = .49$, p

< .001). However, negative affect was not associated with ratings of how sleepy the audience appeared (p > .05) or with cortisol concentrations.

We found no indication that cortisol concentrations and negative affect differed in the VR 200 and IP 2 conditions. Therefore, we concluded that, relative to the VR 2 condition, the effects of the VR 200 condition were more similar to the IP 2 condition. These similarities may be the result of different mechanisms as the larger audience was perceived to be the sleepiest and least attentive audience. It may be that a smaller IP audience elicits cortisol and negative affect due to the clear contingencies of their responses and mere proximity to the performer. A larger VR audience may elicit similar responses due to the sheer number of people apparently rejecting the performer. Future research should focus on these potential, differing mechanisms.

When we considered sex differences, we found that, consistent with previous research (J. J. W. Liu et al., 2017; Q. Liu & Zhang, 2020; Santl et al., 2019), males had significantly more AUCg, AUCi, and average cortisol concentrations than females in all conditions. These differences appear to be driven by differences in reactivity as males and females had similar baseline cortisol concentrations. When examining each of the conditions separately, males had significant reactivity in all conditions and female cortisol reactivity was not statistically significant in any condition. This is consistent with other studies that have found non-statistically significant cortisol responses in females (Kirschbaum et al., 1992). It is notable that, because of this consistent sex difference, previous cortisol reactivity research has tended to enroll only male participants (e.g., Zimmer et al., 2019) limiting our abilities to understand stress-related processes in females.

Although we had no a priori hypotheses concerning the effects of the sex-makeup of audience dyads, analyses of this variable revealed that in the IP condition FF audiences elicited more negative affect than MF audiences. This was not the case in the VR condition. It is possible that in IP conditions FF audiences behaved more critically than MF audiences, leading to differences in negative affect. Pre-recorded VR audiences eliminate this variance by having audiences provide scripted, non-performance contingent responses. Future research should take advantage of the enhanced control afforded by a pre-recorded VR TSST audience and consider further, what variables may be causing between-audience differences in IP TSST protocols.

4.1. Limitations

This study had several limitations. First, we did not include a no-audience condition. Simply wearing an Oculus headset for approximately 15 minutes could potentially affect stress-related variables. Also, without an IP, 200-person audience condition, it is not clear if a large IP audience would elicit a stress response similar to that elicited by the VR 200 pre-recorded audience. Examination of these variables in future research will help us better understand the influence and utility of VR in psychophysiological reactivity studies.

Our protocol included condition differences other than audience size and modality of audience presentation that may have affected stress responses. For example, to enhance the negative valence of the condition, there was a member of the 200-person audience that left the theatre during the performance. Furthermore, to enhance the similarity of the IP and VR audiences, our protocol included the suggestion that the VR audiences could hear and respond to the participants' performances. We did not assess the importance of these factors directly. It is possible that seeing an audience member leave is critical to eliciting stress with a large audience, and that knowledge of an audience being pre-recorded would mitigate stress effects. Lastly, 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 visual acuity lessened the impact of the large audience (Mostajeran et al., 2020).

The use of VR affords many new empirical avenues for the discovery of potentially important moderators of VR TSST effects. For instance, future studies should consider audience sizes between 2 and 200 members, speech preparation with or without paper-and-pencil, and the impact of previous public speaking experience (McKinney et al., 1983) or experience with VR.

Finally, research has indicated that the menstrual cycle of female participants may affect participant responses to the TSST (Childs et al., 2010; Kirschbaum et al., 1999). Not accounting for this variable in the present study may have obscured significant results.

5. Conclusion

Classic IP-TSST protocols are limited by laboratory resources and confounds of audience consistency. This study was the first to consider the impact of a very large, 200-person audience on responses to a TSST protocol using VR technology. Our results added to the growing body of literature suggesting that VR can be used to effectively elicit stress responses. This is the first study to suggest that a VR-TSST may be especially effective when it involves a large 200-person audience. How individuals deal with extreme situations like this is an interesting empirical question. VR technology provides a portable, efficient method for psychological stress research to explore human experiences that would be otherwise unsafe or untenable to emulate in the lab.

6. Funding statement

7. References

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

Bond, R. (2005). Group Size and Conformity. Sage Publications. doi:10.1177/1368430205056464

Bosch, J. A., de Geus, E. J., Carroll, D., Goedhart, A. D., Anane, L. A., van Zanten, J. J., . . . Edwards, K. M. (2009). A general enhancement of autonomic and cortisol responses during social evaluative threat. Psychosom Med, 71(8), 877-885. doi:10.1097/PSY.0b013e3181baef05

Childs, E., Dlugos, A., & De Wit, H. (2010). Cardiovascular, hormonal, and emotional responses to the TSST in relation to sex and menstrual cycle phase. Psychophysiology, 47(3), 550-559. doi:10.1111/j.1469-8986.2009.00961.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. doi:10.1111/j.1467-6494.2004.00295.x

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. doi:10.1037/0033-2909.130.3.355

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. doi:10.1016/j.psyneuen.2021.105626

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. doi:1016/j.psyneuen.2017.02.030

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. Journal of Experimental Psychology: Human Perception and Performance, 35(3), 718-734. doi:10.1037/a0013899

Helminen, E. C., Morton, M. L., Wang, Q., & Felver, J. C. (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. C., Morton, M. L., Wang, Q., & Felver, J. C. (2021). Stress Reactivity to the Trier Social Stress Test in Traditional and Virtual Environments: A Meta-Analytic Comparison. Psychosom Med, 83(3), 200-211. doi:10.1097/PSY.0000000000000918

Hilmert, C. J. (unpublished data). A public speaking self-efficacy questionnaire.

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. doi:10.1089/cpb.2007.9973

Kirschbaum, C., Kudielka, B. M., Gaab, J., Schommer, N. C., & Hellhammer, D. H. (1999). Impact of Gender, Menstrual Cycle Phase, and Oral Contraceptives on the Activity of the Hypothalamus-Pituitary-Adrenal Axis. Psychosomatic Medicine, 61(2), 154-162. doi.org/10.1097/00006842-199903000-00006

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. doi:10.1159/000119004

Kirschbaum, C., Wüst, S., & Hellhammer, D. (1992). Consistent sex differences in cortisol responses to psychological stress. Psychosomatic medicine, 54(6), 648-657. doi:10.1097/00006842-199211000-00004

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. doi:10.1037/0022-3514.45.6.1263

Labuschagne, I., Grace, C., Rendell, P., Terrett, G., & Heinrichs, M. (2019). An introductory guide to conducting the Trier Social Stress Test. Neurosci Biobehav Rev, 107, 686-695. doi:10.1016/j.neubiorev.2019.09.032

Latane, B. (1981). The psychology of social impact. American Psychologist, 36(4), 343-356. doi:10.1037/0003-066X.36.4.343

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. doi:10.3758/BF03208286

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. doi:10.1016/j.beproc.2018.10.003

Liu, J. J. W., 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. doi:https://doi.org/10.1016/j.psyneuen.2017.04.007 doi:10.1016/j.psyneuen.2017.04.007

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

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. doi:10.1111/j.2044-8260.1978.tb00280.x

McEwen, B. S. (1998). Stress, adaptation, and disease. Allostasis and allostatic load. Ann N Y Acad Sci, 840, 33-44. doi:10.1111/j.1749-6632.1998.tb09546.x

McEwen, B. S., & Wingfield, J. C. (2003). The concept of allostasis in biology and biomedicine. Hormones and behavior, 43(1), 2-15. doi:10.1016/S0018-506X(02)00024-7

McKinney, M. E., Gatchel, R. J., & Paulus, P. B. (1983). The Effects of Audience Size on High and Low Speech-Anxious Subjects During an Actual Speaking Task. Basic and Applied Social Psychology, 4(1), 73-87. doi:10.1207/s15324834basp0401_6

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). doi:10.1109/VR46266.2020.00050

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

Pruessner, J. C., Kirschbaum, C., Meinlschmid, G., & Hellhammer, D. H. (2003). Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change. Psychoneuroendocrinology, 28(7), 916-931. doi:10.1016/s0306-4530(02)00108-7

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

Watson, D., & Clark, L. A. (1994). The PANAS-X: Manual for the positive and negative affect schedule-expanded form. doi:10.17077/48vt-m4t2

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