CHENEGA® HEALTHCARE SERVICES, LLC

Assessment of the Psychological and Physiological Effects of Augmented Reality (APPEAR)

Final Technical Report

September 27, 2021

Chenega Healthcare Services, LLC • 5253 Prue Rd, #230 • San Antonio TX 78240

Table of Contents

Assessment of the Psychological and Physiological Effects of Augmented Reality (APPEAR)

Final Report Prepared for U.S. Army Medical Rsearch and Development Command, Fort Detrick, Maryland 21702-5012

1 Introduction and Background

This project was awarded under Medical Technology Enterprise Consortium, Award Number W81XWH1990005, on 8 April 2019, to Chenega Healthcare Services / Regents of the University at Irvine, California. The Principal Investigator was George J. Gisin, Jr., PhD, MHA.

Patient simulators have demonstrated improved learning outcomes in medical training consequently, over the past decade, the use of simulators has become an increasingly important and prominent part of medical training, both civilian and military. For example, the military primarily uses two types of patient simulation in its training programs. These include mechanical manikins (i.e., Laerdal SimMan), and the "buddy" system in which a fellow student pretends to be a patient. These have been shown to be effective in teaching procedural skills such as intubation, surgical procedures and lumbar puncture. However, these methods lack realism and critically do not provoke a realistic emotional response comparable to true emergency medical scenarios in trainees. As such, they may not adequately support development of critical decision-making behaviors in highly emotional contexts. This significantly limits their educational and operational value, as does – in the case of manikins – their cost, reliance on electricity, and lack of portability.

When considering learning as an outcome, it is critical to understand that it is not simply a memorization of facts but rather a complex interplay between cognition, emotional state and physiology. The cognitive, affective and psychomotor components of learning, described in Bloom's taxonomy, are frequently used to provide a framework on which medical curricula are designed. However, there is limited research showing traditional manikin-based simulations influencing the cognitive or affective domains. While this may be due to a lack of studies, we hypothesized it is due, in part, to the fact that manikin-based simulators are not particularly realistic. The solution provided by MedCognition, using a Microsoft Hololens, substantially improves the realism depicted in a variety of clinical scenarios by using an augmented reality (AR) interface. However, it is not known if AR evokes a substantially different response in learners than traditional simulation modalities.

Previous studies have shown the potential for stressors to enhance learning. A stressor is an event, experience, or environmental stimulus that is perceived as a threat or challenge and causes stress in an individual; these can be either physical or psychological. In addition to stress enhancing learning, emotion can play a role. There are multiple proposed mechanisms for improved learning due to emotional stressors including: 1) cortisol peaks in the basolateral amygdala leading to improving memory, 2) emotional stressors during simulation cases "prime the brain for learning during the debriefing process," 3) emotional impact of negative cases motivating students to avoid similar negative outcomes in the future, 4) stress during simulation "inoculates" learners to stress, leading to better performance in real- world stressful situations, and 5) mood congruent memory. Given that three of these five mechanisms involve the emotional response, we believed that evaluating the emotional stress of learners before, after and

during the simulation would be a valuable psychological measure in the effects of AR. There is limited literature on the psychological impact of AR in medical education, with many possible exclusions to participating in such training and whether improved realism and stressors can improve learning.

Due to the potential benefits stress and emotion have on learning, we hypothesized that the higher-fidelity, more realistic AR simulation will more successfully elicit emotional stress compared to a standard medical simulator, which based on previous literature should lead to improved learning and performance. As mechanical manikins are the defacto standard in medical simulation studies, comparing AR to manikins provides a necessary baseline for the psychological assessment.

AR is the use of a computer-based simulation engine to add non-real sensory information to the real sensory world. Essentially, AR directs participants' attention to either existing information that they would have not been consciously aware of or to new information that changes their perceptual information. Although this co-registered information can be visually projected directly onto real objects, AR information is often presented directly to the recipient by a device attached to the recipient.

Overall end goal of program: The ultimate goal of this program was to identify psychological and physiological limitations of AR prototypes currently under development or used for medical simulation training. The research assessed and informed prototype development and/or refinement of existing AR prototypes for medical simulation, reducing overall developmental risk, 1) enabling efficiency in the design of future AR scenarios, 2) identifying potential safety issues, and 3) identifying risk factors for adverse reactions to AR medical simulations. Assessing the physiological and psychological effects of AR prototypes for military medical simulations is imperative to the technological development and refinement needed to fully address the existing capability gaps identified in the Joint Evacuation and Transport Simulation (JETS) and Point of Injury and Trauma Simulation (POINTS) programs and deliver effective solutions to the Warfighter.

1.1 Procedure

The Team consisted of personnel from Chenega Healthcare Servcies (CHS), and the University of California Irvine (UCI). The CHS personnel included Grace Castillo, General Manager, George Gisin, PhD, MHA, and Terrisa Ducate, Program Manager. The UCI Team consisted of Shannon Toohey, MD, MAEd, Sarah Pressman, PhD, MS, and Alisa Wray, MD, MAEd.

The Team held a Kickoff Meeting on 20 May 2019 and during the next two months, the Team achieved several accomplishments. We decided a within-subjects experimental design would be used to assess the psychological and physiological responses of the students. We had planned to use military personnel; however, this was impractical because the approval process to use active-duty personnel is very lengthy far exceeding the time allocated for the Period of Performance. We thus decided to use a population of 2nd year medical students who have almost no clinical experience, comparable to most military, non-medical personnel.

We determined the facilities and student body at the University of California, Irvine were sufficient to complete the study and provide the investigators access to a simulation laboratory equipped with mechanical manikins, obviating the need to purchase and house expensive mechanical manikins. The study we proposed had the students complete similar medical

simulation scenarios on both a manikin based simulator, such as a SimMan™, and on the MedCognition augmented reality system, PerSim™, while measuring psychological parameters and evidence of stress. The crossover design would allow for comparison of each student's psychologic response and minimizes confounding due to variance in the individual psychological responses, as students would act as their own controls. Students would complete two scenarios centered on pediatric resuscitation and subsequent death of the patient: one status asthmaticus and one pediatric sepsis, both with unstable vital signs requiring acute resuscitation who would ultimately succumb to their illness regardless of learner actions. These cases would be integrated into the medical student curriculum with the objective of covering personal emotional stressors in work and difficult conversations but would allow us to maximize the specific psychologic effects we wanted to evaluate while being broadly applicable to a variety of care settings and learners.

To assess the affective responses to these scenarios, we planned to measure emotion change using a battery of validated and widely used questionnaires (e.g., items drawn from the Circumplex Model of Affect which assesses both high and low arousal emotional experiences). Specifically, we would use the State Affect Questionnaire (StAQ). In addition, we would state emotion changes in other relevant emotions like experiences of stress, frustration, engagement and boredom using items drawn from existing state assessments, the Positive and Negative Affect Schedule (PANAS). The primary goal of these analyses was to see whether there were significant differences in the emotional changes from baseline when utilizing AR versus mannequins. On the flip side, we would also look for possible emotional benefits of utilizing AR. Are participants more engaged, more interested, and less bored? Are they more excited by the learning experience? We would explore this utilizing the same mood scales discussed above both immediately after the experience and 30 minutes later.

We would also do more exploratory analyses via debriefing interviews following these learning experiences. These interviews would be coded by psychological researchers using Linguistic Inquiry and Word Count (LIWC), a validated text analysis program that is widely used in psychologic research. We concluded this analysis would provide a more nuanced examination of the cognitive and emotional experience of using AR versus older methods and provide convergent data to the self-reported emotions analysis. Next, while less relevant for learning, it would be important to examine wellbeing outcomes related to using AR since this had never been examined in this context. We planned to examine the emotional experiences of individuals immediately following the simulation (i.e. how did they feel during the experience) as well as 30 min later and the following week. This would allow us to see if there were lasting negative emotional effects (e.g., lasting depression, anxiety) after the learning experience was over and whether this was a risk that should be recognized in this approach. While we did not expect serious problems in the average participant, we planned to also examine interactions between the simulation and baseline psychological disorder (e.g., presence of clinical depression, anxiety disorder, and past trauma using standard self-report assessments) and whether individuals entering with poor mental health were adversely impacted by this experience. Correlation of the debriefing abstraction and mood and emotion questionnaires could help identify signs of individuals who may have negative or traumatic reaction to AR simulation who may need additional support or resources to help them adapt to that stimulus. We would provide any needed resources to those individuals after the simulation sessions.

Additional secondary measures would include salivary cortisol levels (SC), which have previously been shown to have high correlation with psychologic stress and were considered an objective indicator of the psychological stress response. In addition, we would monitor electrodermal activity (EDA) as this has previously been shown to be a marker for emotional experience and arousal. Adverse side effects would be measured with the Simulator Sickness Questionnaire, a validated measurement for simulation side effects that have been previously reported in virtual reality literature. However, because of a completely different approach to display technology, we anticipated that the augmented reality system in this experiment would less likely evoke these symptoms. Lastly, we planned to record audio and video during the sessions and review them for evidence of collisions or trips and falls to identify potential physical safety issues caused by the AR technology.

The investigators at the University of California, Irvine collectively had decades of experience in education, simulation and psychologic research. Furthermore, they had access to the necessary simulation center and learners necessary to complete the proposed study. Additionally, they have experience writing simulation cases, and have been teaching with both the MedCognition PerSim™ simulator and Laerdal SimMan™ manikin simulator for over three semesters. Additionally, the larger team headed by Chenega Corporation and assisted by MedCognition, had key personnel with extensive experience in the Army Medical Department. As such, they had already made inquiries with military medical organizations regarding this project and were uniquely qualified to help merge the efforts of the UCI researchers and any military organizations that may participate.

The team then developed study protocol and Institutional Review Board (IRB) applications to be submitted to the University of California Irvine, Irvine, CA (UCI) Institutional Review Board. Since our research would include human subjects, it was mandatory to have approval from the IRB. Following this approval, it would be essential to request approval from the Human Research Protection Office (HRPO), U.S. Army Medical Research and Development Command (USAMRDC).. This additional approval was required since the research was funded by the U.S. Army.

While we waited for approval from the UCI IRB, the team prepared didactic and simulation material for the pediatric resuscitation cases. UCI also acquired simulation systems using the Microsoft Hololens from MedCognition. They hired personnel to perform the study, plus clean and analyze the data. They also acquired electrodermal activity monitors and software needed to collect data on the cardiovascular signs of psychologic stress.

On 30 October 2019, we received initial review from the IRB UCI, requiring minor edits. Within a few weeks, final approval was provided by the IRB. We immediately requested approval from HRPO USAMRDC, and after several edits, we received approval on 19 March 2020. While waiting for the approval, the Team completed testing on the simulation didactic materials and equipment.

Finally, on 26 March 2020, we commenced research on the medical students and on 29 September 2020, we completed the study on a total of 86 total participants who completed all study parts. This was 14 more subjects than the 72 subjects required for statistical significance.

A copy of the raw study was provided to the government. On 2 December 2020, the Team began cleaning, coding, and analyzing the data. They completed this effort on 31 March 2021. On 30

April 2021, our Team completed the data summary, and provided a manuscript for submission to an academic journal.

A brief discussion of the findings will now be provided. Following the discussion is a detailed summary of the entire study.

1.2 Findings

Both the manikin and AR simulators elicited emotional (i.e., a reduction in positive emotion and an increase in negative emotion) and stress responses during and after the simulations. This is consistent with previous studies that showed that simulation in medical education can elicit a stress response as well as a range of emotional and cognitive changes. Given previous research showing stressors can enhance learning, this suggests both modalities of simulation can have a beneficial effect for learners, although future studies will need to evaluate actual learning outcomes. Of note, there was some concern that AR might be associated with a dangerously high level of stress because of the added realism and interactive nature, however, it does not seem to be any more stressful than past medical training approaches adding some indication that this is not a concern at least in this context. Further sub-analysis of participants with pre-existing PTSD, perceived stress and depression did not show statistically significant differences in stress with AR simulation, suggesting that even those with pre-existing conditions do not need to be excluded from AR technology. Further, the levels of stress and negative emotion reported in these simulations do not appear to be at levels that are different than other study averages.

Augmented reality technology is relatively new and its ability to elicit a stress response when compared to standard manikin simulation technology could help guide future educational practices and research. Our study shows no significant differences between AR and standard manikin simulation technology, except a small difference where the increase in skin conductance in response to the manikin simulator was significantly higher than that of AR. Cortisol differences, however, were not different across the two platforms. This suggests that the AR approach is psychologically and physically comparable to standard manikin-based simulators, and perhaps even less physiologically arousing than past learning modalities.

Given the nature of the simulations involving pediatric deaths it is not surprising that the overall stress increased during and after each simulation. However, students show decreased stress levels in their second simulation. Previous studies have shown that stress factors in simulation-based training may help with the acquisition of stress management skills. In addition to stress management skills, it could suggest a desensitization to the simulation regardless of type of simulator. Chang et al. suggested that VR simulation could be used to desensitize pediatric physicians from stressful situations based on his study evaluating VR stress response and real-life situations. Although, Hardernberg, et al. showed no decreased stress response in nursing students with repeated simulations. Our results contradict this, which could be very useful for educational efforts and future training for many types of medical practitioners who experience high stress situations.

1.3 Limitations

This is a single site study comparing AR simulation to standard manikin-based simulation. While we attempted to look at multiple evaluators of emotional stress (cortisol, self-reported stress, electrodermal activity) these still may not have fully captured the stress response of the students. Furthermore, in an attempt to identify differences in stress response we intentionally used a very high stress case. Other simulation cases that are not as high-stakes likely elicit lower stress levels.

Furthermore, cortisol and electrodermal activity have confounding factors such as gender, medications, and time since waking. While corrected for these confounders there may be others unaccounted for that resulted in noise in the data.

1.4 Conclusions & Implications

Augmented reality simulators elicited similar stress responses to manikin-based simulators suggesting they are comparable tools for medical education. Furthermore, there was no evidence of AR simulators causing excessive stress to participants. Future research should evaluate whether AR simulators increase learning outcomes or help with desensitization or stress management skills with repeated use.

2 Summary of Study

The study sample consisted of second-year medical students (N=89) at the University of California, Irvine. Students were evaluated while completing both AR and standard medical simulation cases on mechanical manikins as part of their training. There were no exclusion criteria and any student who wanted to participate was eligible.

The objective of the Program was to complete a human study that compared psychologic effects, specifically those representing an emotional stress reaction, of an AR simulator to that of a standard medical simulator (SimMan) during medical education training. We believed having the manikin control group is essential when assessing the psychologic effects of AR simulation as it allowed effects to be compared to the current standard in simulation training. The study assessed the following Primary and Negative Outcomes:

1. Primary outcome:
    i. Negative emotional responses and changes is stress, as assessed by validated mood and emotion questionnaire. Specifically, we measured before and after state affect (State Affect Questionnaire [StAQ] and Positive and Negative Affect Schedule [PANAS]) and examine to what extent the change scores in state (in the moment) negative and positive emotions differ across conditions (AR vs. mannequin).
2. Secondary outcomes for psychologic assessment:
    i. Qualitative analysis of the debriefing discussions for emotional words, cognitive processing words and types of emotion.
    ii. Median salivary cortisol levels over the course of the procedures, which has previously been shown to have high correlation with psychologic stress. Specifically, we measured changes from pre-simulation to immediately after the simulation and 15 minutes after the simulation was over, when cortisol levels typically peak (i.e., to assess maximal stress).
    iii. Electrodermal activity (EDA), as this has previously been shown to be a marker for emotional experience.
    iv. Simulator Sickness Questionnaire, a validated measure of adverse side effects during virtual reality simulation.
    v. Rate of collisions or trip and falls to identify potential physical safety issues caused by the AR technology.

3 Requirements:

3.1 Task 1: Human Subjects Study Planning

The Team complied with restrictions and reporting requirements for the use of human subjects, to include research involving the secondary use of human biospecimens and/or human data. The team also had agreements in place for intellectual property and Microsoft Hololens at no additional cost to the Government.

The Team finalized the study protocol for the evaluation of the psychological effects of AR in medical simulation training. The team then prepared and piloted didactic and simulation content on pediatric resuscitations to be used to assess the psychological effects of AR in medical simulation training.

As stated earlier, the team obtained approval from the UCI Institutional Review Board (IRB); and, then provided this approval to the U.S. Army Human Research Protections Office (HRPO), that subsequently provided approval.

3.2 Task 2: Human Study Assessment Execution

With approval of the UCI IRB and the HRPO, the CHS Team commenced hiring personnel to perform the study and clean and analyze the data. The Team also acquired necessary simulation systems (MedCognition Persim systems) to run the AR medical simulation training. Finally, the team acquired electrodermal activity monitors necessary to collect data on cardiovascular signs of psychologic stress.

3.2.1 Procedures

Medical students completed similar medical simulation scenarios, three weeks apart, on both a manikin-based simulator, SimMan™, and on the MedCognition AR system, PerSim™, while measuring psychological parameters and evidence of stress. The within-subjects design allowed for comparison of each student's psychologic response and minimizes confounding due to variance in the individual psychological responses, as students acted as their own controls. Before participating in the study sessions, participants completed a baseline questionnaire from home which assessed health behaviors, trait emotion, and demographic characteristics relevant for controls. Within each study session, students completed one-of-two scenarios centered on pediatric resuscitation and subsequent death of the patient: one status asthmaticus and one pediatric sepsis, both with unstable vital signs requiring acute resuscitation who ultimately succumbed to their illness regardless of learner actions. These cases were integrated into the medical student curriculum with the objective of covering personal emotional stressors in work and difficult conversations but also allowed maximum specific psychological effects. The scenarios lasted approximately 10 minutes each.

Electrodermal activity was continuously monitored for before, during, and after the scenario to establish baseline and recovery periods. These sections were utilized to collect baseline, reactivity, and recovery data for galvanic skin response. Additionally, salivary cortisol samples were collected to align with times before, immediately following, and 15 minutes after each simulated case. Psychological data was collected through surveys administered before and immediately following each simulation session. These psychological surveys contained questionnaires measuring state stress and affect. The post-simulation survey additionally included qualitative debriefing questions related to the passing of the participant and the medical knowledge of the participant. Throughout the week following each of the two simulations,

participants were instructed to fill out a questionnaire from home that measured psychological variables relevant to prolonged stress and mental well-being.

3.2.2 Measures

Psychological Stress

The current study compares psychological effects, specifically those representing an emotional stress reaction, of an AR simulator to that of a standard medical simulator (SimMan) during medical education training. To measure perceived stress responses induced from the simulation, slider scales ranging from 1-100 were utilized to capture stress levels before and after the simulation.

Positive and Negative Emotion

State Affect

To assess the affective responses to these scenarios, we measured emotion change using the State Affect Questionnaire (StAQ). In addition, we investigated state emotion changes in emotions relevant to positive and negative emotions using items drawn from existing state assessments, the positive and negative affect schedule (PANAS) [15]. Positive and negative affect subscales within the PANAS were utilized to create variables for positive and negative affect. Mean scores were then calculated for positive and negative affect by utilizing subscales within the PANAS that yield a positive and negative affect score respectively.

Qualitative Debriefing Survey

Additionally, we conducted more exploratory analyses on positive and negative emotion via open-answer debriefing surveys following the simulation experiences. These surveys were coded by a psychological researcher using Linguistic Inquiry and Word Count (LIWC), a validated text analysis program that is widely used in psychological research. A positive affect variable was created by utilizing an affect dictionary within the LIWC. The dictionary analyses passages for percentages of mood-oriented words that suggest positive or negative emotion.

This analysis provides a more nuanced examination of the cognitive and emotional experience of using AR versus older methods and provides convergent data to the self-reported emotions analysis.

Wellbeing Outcomes

Next, while less relevant for learning, it was important to examine wellbeing outcomes related to using AR since it has never been examined in this context. We examined the emotional experiences of individuals immediately following the simulation (i.e. how did they feel during the experience) as well as 30 minutes later and the following week. This allowed us to see if there are lasting negative emotional effects (e.g., lasting depression, anxiety) after the learning experience is over and whether this is a risk that should be recognized in this approach. While we did not expect serious problems in the average participant, we also examined interactions between the simulation and baseline psychological disorder (e.g., presence of clinical depression, anxiety disorder, and past trauma using standard self-report assessments) and whether individuals who entered with poor mental health are adversely impacted by this experience. Correlation of the debriefing abstraction and mood and emotion questionnaires can help identify signs of individuals who may have negative or traumatic reactions to AR simulation who may need

additional support or resources to help them adapt to that stimulus. This is essential to know, as we can provide any needed resources to these individuals after the simulation sessions.

Physiological Stress

Salivary Cortisol

Additional secondary measures included salivary cortisol levels, which has previously been shown to have high correlation with psychological stress and is considered an objective indicator of the psychological stress response. Median salivary cortisol levels were measured over the course of the procedures, which has previously been shown to have high correlation with psychological stress. Specifically, we will measure changes from pre-simulation to immediately after the simulation and 15 minutes after the simulation is over, when cortisol levels typically peak (i.e., to assess maximal stress). Sample size was calculated based on a similar previous study and their median salivary cortisol level differences. The crossover study design allows learners to act as their own controls and decreases costs and total number of subjects needed.

Experimental sessions were conducted in the afternoon (between 12:00-5:00pm) to account for the diurnal rhythm of cortisol. Salivettes were stored at -80°C until batch analysis at the end of data collection at the laboratory of the Institute for Interdisciplinary Salivary Bioscience Research (University of California Irvine, Irvine, CA). Before assaying, the samples were thawed for an hour to return to room temperature. For cortisol, all samples were assayed in duplicate using an expanded range high sensitivity salivary cortisol enzyme immunoassay kit (Salimetrics, LLC; State College, PA). The assay range of sensitivity was 0.007 to 3.0 ug/dl and the average intra-assay coefficient of variation was 5.5%.

Electrodermal Activity

In addition to salivary cortisol levels, we monitored galvanic skin response (GSR) as this has previously been shown to be a marker for emotional experience and arousal [20]. Internal and external stimuli can activate the sweat glands, which allows electrical current to flow, increasing GSR. This response to stimuli can act as a biomarker for stress and emotional responses. The GSR data was collected via a small unobtrusive device (Shimmer3) that was monitored by the researchers. The device was placed on a wristband that was fastened to participants' wrists. To collect GSR data, the device had two wires that extended from the hardware and was attached to participants' palms with two electrodes.

Researchers monitored the GSR data utilizing Bluetooth connectivity through a laptop. The researchers took notes of any artifacts that could cause spikes in the data unrelated to the simulation, such as coughing, external noises, etc. This is essential to monitor for, as auditory arousal has been linked to higher heart rate and activation of the sweat glands, effecting GSR data [21]. Additionally, researchers made note of any participants that had connectivity issues due to exceptionally sweaty palms. GSR means were utilized in analyses by obtaining the average GSR score for the baseline, reactivity, and recovery phases of each simulation session.

Simulation Side Effects

Adverse side effects were measured with the Simulator Sickness Questionnaire, a validated measurement for simulation side effects that have been previously reported in virtual reality literature. However, because of a completely different approach to display technology, we anticipated that the AR system in this experiment is less likely to evoke these symptoms.

Analytic Strategy

Linear Mixed Model (LMM) for repeated measurements was used for data analysis by using the MIXED command in SPSS Statistics software (IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp). Simulation type and time of measurements were considered as fixed effect variables and the participants as random effect variable. A separate LMM analysis was performed for each dependent variable, adjusting for potential confounders accordingly. The correlation between repeated measurements within subjects was considered as "unstructured". A Square root (SQRT) transformation was applied to Mean GSR and SSQ, and Natural Logarithm (Ln) transformation was applied to Cortisol before LMM analysis. A p-value < 0.05 was considered statistically significant.

To examine whether Perceived Stress, Depression and PTSD modify the effect of AR on cortisol and Galvanic Skin Response, a LMM analysis was applied to AR data only by including the potential effect modifiers. If the p-value of a potential effect modifier was greater than 0.05, its effect modification on association between AR and the dependent variables was excluded.

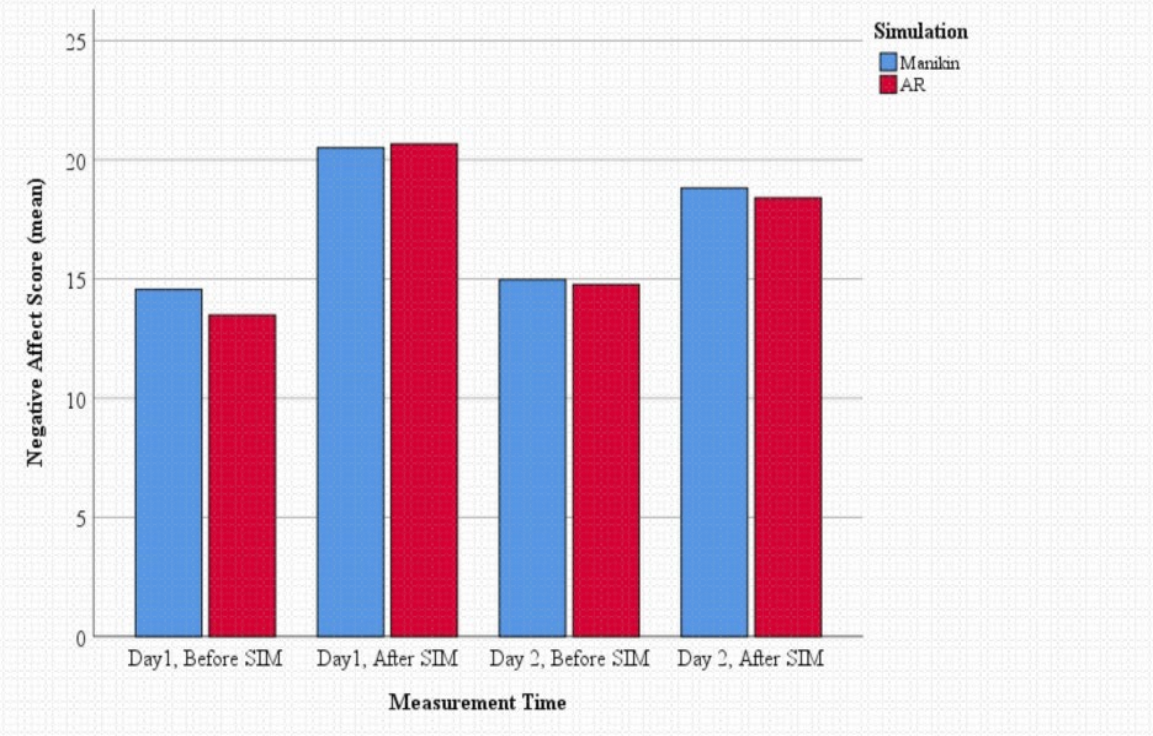
Results

Thirty-seven (43.0%) male and 49 (57.0%) female second-year medical students with a mean age of 25.2 (SD=2.09, 22-30) and 24.7 (SD=2.08, 23-36) respectively, participated in the study from March 2020 to October 2020.

I. Main Effects

I.a Negative Affect

Negative affect showed an increase with either simulations ($p < .001$) but the difference between simulation types was not statistically significant, adjusted for the day of experiment.



Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	p-value	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	19.01	0.71	112.07	26.75	<0.001	17.60	20.42
[Simulation=0]	-5.08	0.57	135.03	-8.96	<0.001	-6.21	-3.96
[Simulation=1]	-0.41	0.56	75.87	-0.72	0.475	-1.53	0.72
[Simulation=2]	0.00[b]	0.00	.	.	.	.	.
[Day=1]	-0.14	0.35	81.68	-0.40	0.687	-0.84	0.56
[Day=2]	0.00[b]	0.00	.	.	.	.	.

a. Dependent Variable: Negative Affect Score.
b. This parameter is set to zero because it is redundant.

Pairwise Comparisons

(I) Simulation	(J) Simulation	Mean Difference (I-J)	Std. Error	df	p-value	95% Confidence Interval for Difference	
						Lower Bound	Upper Bound
None	Manikin	-4.68*	0.56	131.67	<0.001	-5.79	-3.57
	AR	-5.08*	0.57	135.03	<0.001	-6.21	-3.96
Manikin	None	4.68*	0.56	131.67	<0.001	3.57	5.79
	AR	-0.41	0.56	75.87	0.475	-1.53	0.72
AR	None	5.08*	0.57	135.03	0.000	3.96	6.21
	Manikin	0.41	0.56	75.87	0.475	-0.72	1.53

Based on estimated marginal means

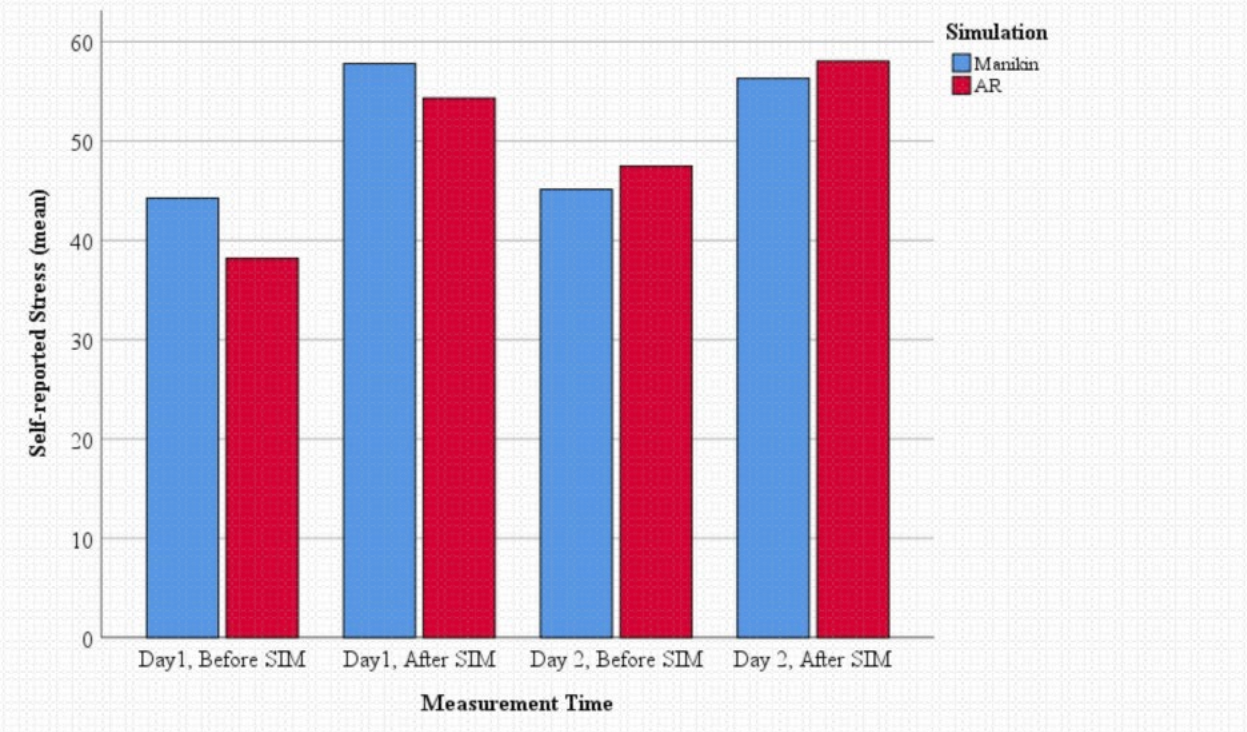
*. The mean difference is significant at the .05 level.

a. Dependent Variable: Negative Affect Score.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

I.b Self-reported Stress

Self-reported Stress showed an increase with either simulations ($p < .001$) but the difference between simulation types was not statistically significant, adjusted for day of experiment and sex of participants. Self-reported stress was lower on the second day overall ($p=.04$).



Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	p-value	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	58.73	3.15	109.51	18.67	<0.001	52.50	64.97
[Simulation=0]	-12.75	1.38	140.83	-9.26	<0.001	-15.47	-10.03
[Simulation=1]	-0.53	1.45	82.09	-0.37	0.71	-3.42	2.35
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[Day=1]	-3.54	1.71	82.00	-2.07	0.04	-6.93	-0.14
[Day=2]	0.00b	0.00	.	.	.	.	.
[sex=1]	-2.40	4.19	83.69	-0.57	0.57	-10.73	5.92
[sex=2]	0.00b	0.00	.	.	.	.	.

a. Dependent Variable: Self-reported Stress.
b. This parameter is set to zero because it is redundant.

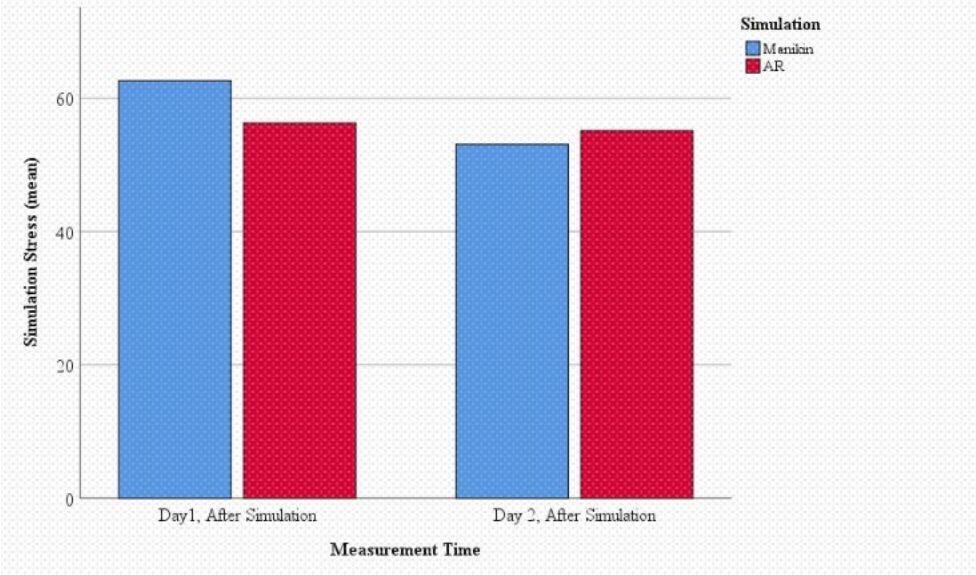
Pairwise Comparisons

(I) Simulation	(J) Simulation	Mean Difference (I-J)	Std. Error	df	p-value	95% Confidence Interval for Difference Lower Bound	95% Confidence Interval for Difference Upper Bound
None	Manikin	-12.21*	1.36	136.63	<0.001	-14.90	-9.53
	AR	-12.75*	1.38	140.83	<0.001	-15.47	-10.03
Manikin	None	12.21*	1.36	136.63	<0.001	9.53	14.90
	AR	-0.53	1.45	82.09	0.71	-3.42	2.35
AR	None	12.75*	1.38	140.83	<0.001	10.03	15.47
	Manikin	0.53	1.45	82.09	0.71	-2.35	3.42

Based on estimated marginal means
*. The mean difference is significant at the 0.05 level.
a. Dependent Variable: Self-reported Stress.
b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

I.c Simulation Stress

Simulation stress was higher in day 1 compared to day 2 (p= 0.030); however, the difference between the simulation types was not statistically significant (p=0.367), adjusted for day of experiment and sex of the participants.



Estimates of Fixed Effects

						95% Confidence Interval	
Parameter	**Estimate**	**Std. Error**	**df**	**t**	**p-value**	**Lower Bound**	**Upper Bound**
Intercept	55.07	3.34	123.34	16.48	<0.001	48.46	61.69
[Day=1]	5.29	2.40	82.45	2.21	0.030	0.52	10.06
[Day=2]	0.00[b]	0.00	.	.	.	.	.
[Simulation=1]	2.17	2.39	82.75	0.91	0.367	-2.59	6.94
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[sex=1]	-4.82	4.24	84.00	-1.14	0.259	-13.26	3.61
[sex=2]	0.00[b]	0.00	.	.	.	.	.

a. Dependent Variable: Simulation Stress.
b. This parameter is set to zero because it is redundant.

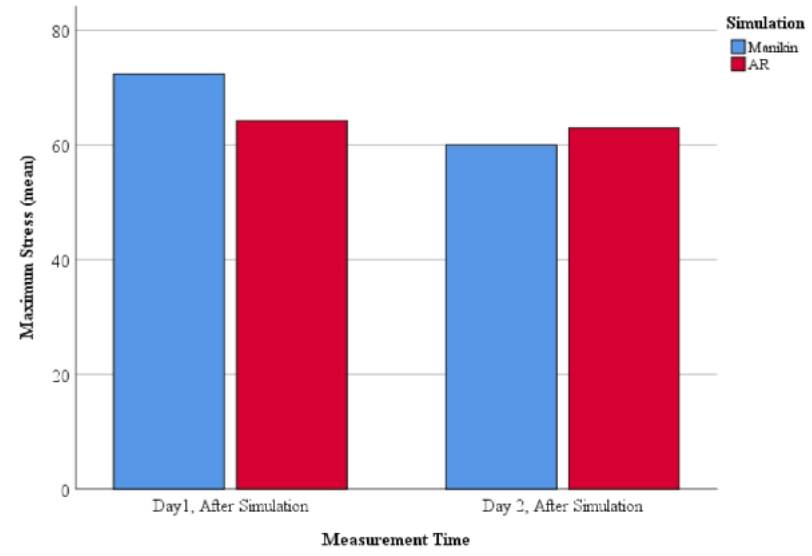
Pairwise Comparisons

						95% Confidence Interval	
(I) Simulation	(J) Simulation	Difference (I-J)	Std. Error	df	p-value	Lower Bound	Upper Bound
AR	Manikin	-2.17	2.39	82.75	0.367	-6.94	2.59

Based on estimated marginal means
a. Dependent Variable: Simulation Stress.
b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

I.d Maximum Stress

Maximum stress was higher in the first day compared to the second day (p=0.001); however, there was no statistically significant difference between the two simulation types in terms of Maximum stress (p=0.120), adjusted for day of experiment and sex of the participants.



Estimates of Fixed Effects

						95% Confidence Interval	
Parameter	**Estimate**	**Std. Error**	**df**	**t**	**p-value**	**Lower Bound**	**Upper Bound**
Intercept	61.94	3.38	115.22	18.32	<0.001	55.24	68.64
[Day=1]	6.60	1.95	81.35	3.38	0.001	2.72	10.49
[Day=2]	0.00b	0.00	.	.	.	.	.
[Simulation=1]	3.02	1.92	82.15	1.57	0.120	-0.80	6.83
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[sex=1]	-4.09	4.44	83.88	-0.92	0.359	-12.92	4.74
[sex=2]	0.00b	0.00	.	.	.	.	.

a. Dependent Variable: Maximum Stress.
b. This parameter is set to zero because it is redundant.

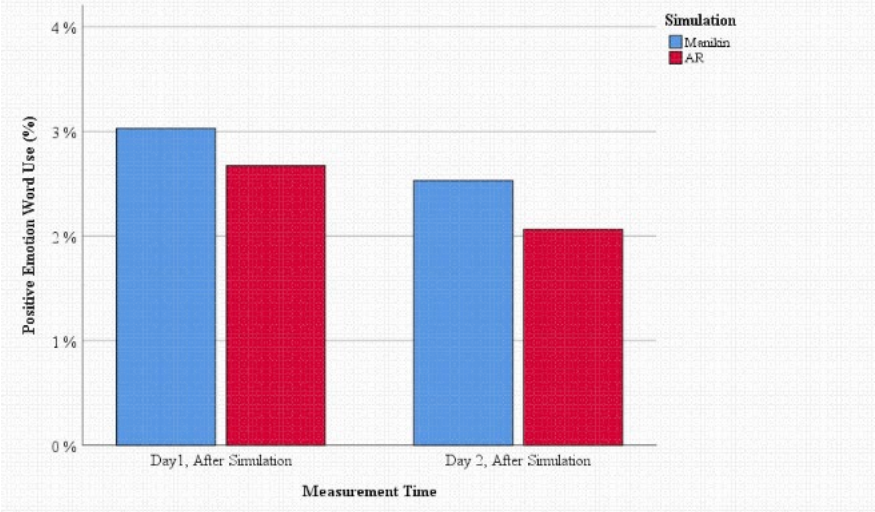
Pairwise Comparisons

						95% Confidence Interval	
(I) Simulation	(J) Simulation	Difference (I-J)	Std. Error	df	p-value	Lower Bound	Upper Bound
AR	Manikin	-3.02	1.92	82.15	0.120	-6.83	0.80

Based on estimated marginal means
a. Dependent Variable: Maximum Stress.
b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

I.e Positive Emotion Words Use Percentage

The percentage of positive emotion words use was higher in the first day of experiment (p=0.021). There was no statistically significant difference between the simulation types in terms of percentage positive emotion words use (p=0.118), adjusted for day of experiment and sex of the participants and the word count for LLWC.



Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	p-value	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	61.94	3.38	115.22	18.32	<0.001	55.24	68.64
[Day=1]	6.60	1.95	81.35	3.38	0.001	2.72	10.49
[Day=2]	0.00b	0.00	.	.	.	.	.
[Simulation=1]	3.02	1.92	82.15	1.57	0.120	-0.80	6.83
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[sex=1]	-4.09	4.44	83.88	-0.92	0.359	-12.92	4.74
[sex=2]	0.00b	0.00	.	.	.	.	.
WordCount_for_LIWC	0.00	0.00	121.02	-0.576	0.571	-0.01	0.00

a. Dependent Variable: Positive Emotion Word Use percentage.
b. This parameter is set to zero because it is redundant.

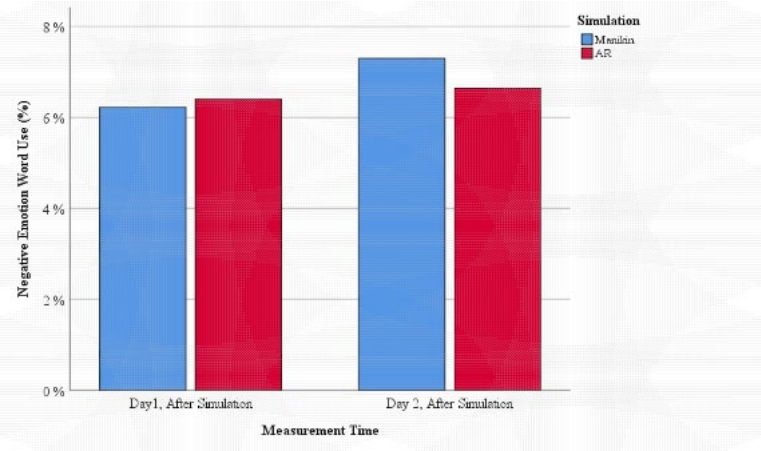
Pairwise Comparisons

(I) Simulation	(J) Simulation	Difference (I-J)	Std. Error	df	p-value	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
AR	Manikin	-0.40	0.25	83.35	0.118	-0.91	0.10

Based on estimated marginal means
a. Dependent Variable: Positive Emotion Word Use percentage.
b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments)..

I.f Negative Emotion Word Use Percentage

There was not a statistically significant difference between simulation types in terms of the percentage of negative emotion word use (p=0.406) adjusted for day of experiment, sex and the word count for LIWC).



Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	p-value	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	9.20	0.60	155.40	15.34	<0.001	8.01	10.38
[Day=1]	0.13	0.43	94.98	0.30	0.764	-0.72	0.98
[Day=2]	0.00b	0.00	.	.	.	.	.
[Simulation=1]	0.33	0.40	81.54	0.84	0.406	-0.46	1.12
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[sex=1]	-0.83	0.52	81.92	-1.59	0.116	-1.86	0.21
[sex=2]	0.00b	0.00	.	.	.	.	.
WordCount_for_LIWC	-0.02	0.00	113.59	-5.29	<0.001	-0.02	-0.01

a. Dependent Variable: Negative Emotion Word Uses (%).
b. This parameter is set to zero because it is redundant.

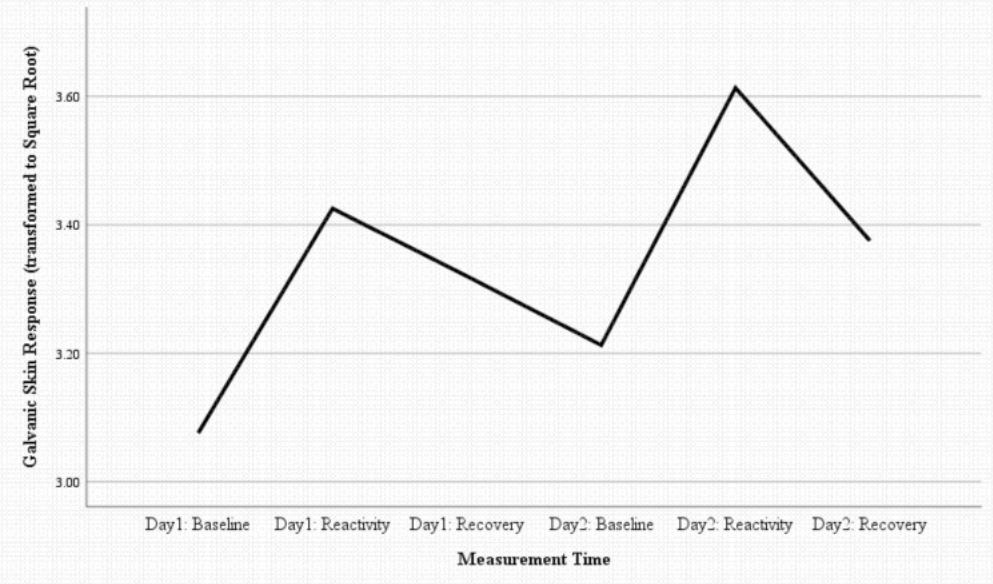
Pairwise Comparisons

(I) Simulation	(J) Simulation	Mean Difference (I-J)	Std. Error	df	p-value	95% Confidence Interval for Difference Lower Bound	95% Confidence Interval for Difference Upper Bound
AR	Manikin	-0.33	0.40	81.54	0.406	-1.12	0.46

Based on estimated marginal means
a. Dependent Variable: Negative Emotion Word Uses (%).
b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

I.g Galvanic Skin Response

Both simulations were associated with increased galvanic skin response ($p< 0.001$). Galvanic skin response was higher in manikin group ($p=0.009$) adjusted for day, sex and medications taken by the participants.



Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	p-value	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	3.43	0.13	99.80	26.60	<0.001	3.17	3.68
[Day=1]	-0.11	0.11	79.03	-0.98	0.332	-0.33	0.11
[Day=2]	0.00b	0.00	.	.	.	.	.
[Simulation=1]	-0.28	0.04	119.66	-7.36	<0.001	-0.35	-0.20
[Simulation=2]	0.11	0.04	78.29	2.69	0.009	0.03	0.18
[sex=1]	0.00b	0.00	.	.	.	.	.
[sex=2]	0.29	0.16	81.48	1.84	0.069	-0.02	0.61
[medication=1]	0.00b	0.00	.	.	.	.	.
[medication=2]	0.16	0.19	83.25	0.84	0.405	-0.22	0.54

a. Dependent Variable: Galvanic Skin Response (transformed into Square Root).
b. This parameter is set to zero because it is redundant

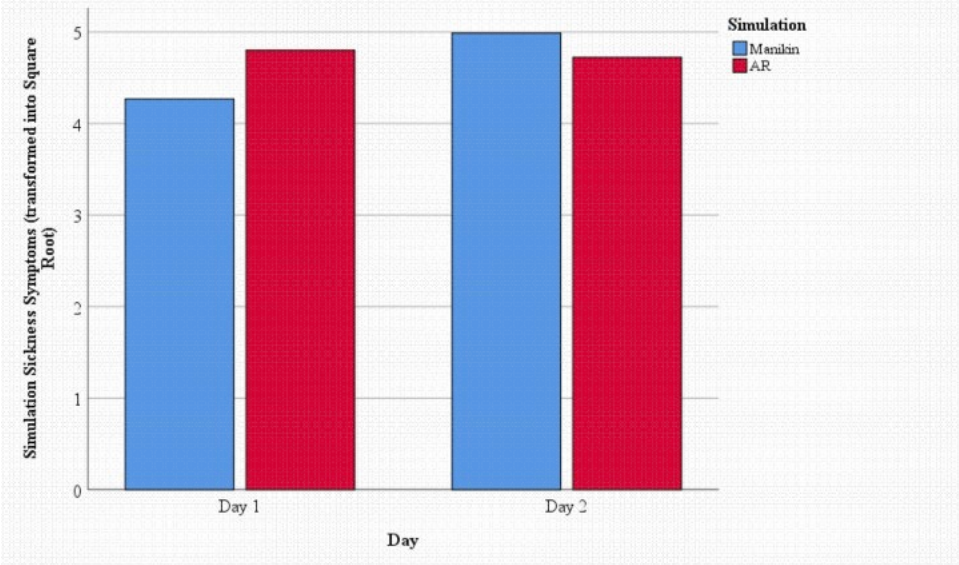
Pairwise Comparison

(I) Simulation	(J) Simulation	Mean Difference (I-J)	Std. Error	df	p-value	95% Confidence Interval for Difference	
						Lower Bound	Upper Bound
None	Manikin	-0.38*	0.04	126.44	<0.001	-0.46	-0.31
	AR	-0.28*	0.04	119.66	<0.001	-0.35	-0.20
Manikin	None	0.38*	0.04	126.44	<0.001	0.31	0.46
	AR	0.11*	0.04	78.29	0.009	0.03	0.18
AR	None	0.28*	0.04	119.66	<0.001	0.20	0.35
	Manikin	-0.11*	0.04	78.29	0.009	-0.18	-0.03

Based on estimated marginal means
*. The mean difference is significant at the 0.05 level.
a. Dependent Variable: Galvanic Skin Response (transformed into Square Root).
b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

I.h Simulation Sickness Symptoms Score

There was not a statistically significant difference in simulation sickness symptoms score between the simulation groups (p= 0.469), adjusted for day of experiment and sex of the participants.



Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	p-value	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	5.14	0.34	121.68	14.91	<0.001	4.46	5.82
[Day=1]	-0.17	0.23	82.70	-0.73	0.469	-0.62	0.29
[Day=2]	0.00b	0.00	.	.	.	.	.
[Simulation=1]	-0.35	0.23	82.06	-1.50	0.137	-0.80	0.11
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[sex=1]	-0.41	0.44	84.09	-0.93	0.357	-1.28	0.47
[sex=2]	0.00b	0.00	.	.	.	.	.

a. Dependent Variable: Simulation Sickness Symptom Score (transformed into Square Root).
b. This parameter is set to zero because it is redundant.

Pairwise Comparisons

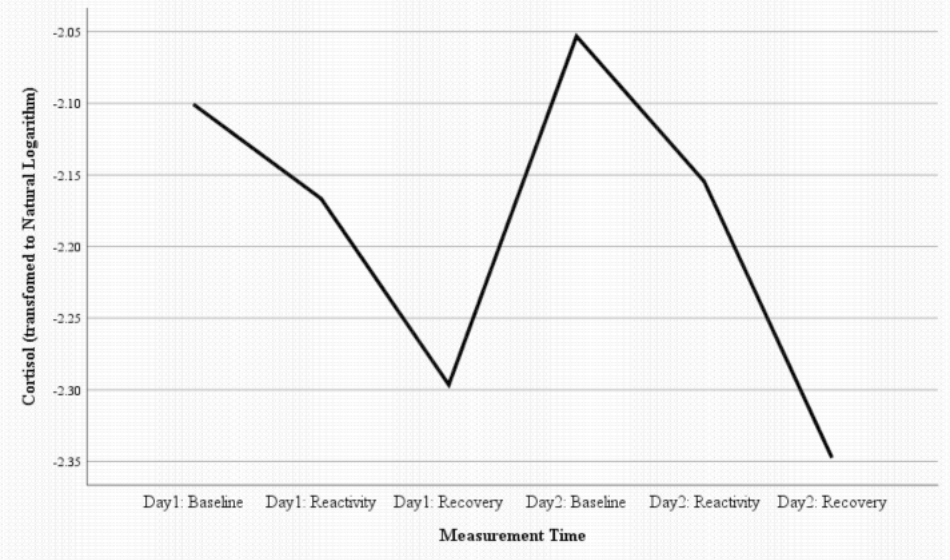
(I) Simulation	(J) Simulation	Mean Difference (I-J)	Std. Error	df	p-value	95% Confidence Interval for Difference Lower Bound	95% Confidence Interval for Difference Upper Bound
AR	Manikin	0.17	0.23	82.70	0.469	-0.29	0.62

Based on estimated marginal means
a. Dependent Variable: Simulation Sickness Symptom Score (transformed into Square Root).
b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

I.i Cortisol

There was not a statistically significant difference in mean cortisol level between the simulation groups (p= 0.409), adjusted for the day of experiment, sex of the participants, use of medication

by the participants, and the time past from wakeup to simulation. Overall, cortisol was higher in male participants (p= 0.023).



Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	p-value	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	-1.88	0.19	147.17	-10.05	<0.001	-2.25	-1.51
[Day=1]	-0.05	0.06	75.12	-0.85	0.400	-0.17	0.07
[Day=2]	0.00b	0.00	.	.	.	.	.
[Simulation=0]	0.06	0.04	134.67	1.56	0.122	-0.02	0.14
[Simulation=1]	-0.04	0.05	74.07	-0.83	0.409	-0.13	0.05
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[sex=1]	-0.41	0.44	84.09	-0.93	0.357	-1.28	0.47
[sex=2]	0.00b	0.00	.	.	.	.	.
[Medication=1]	0.00b	0.00	.	.	.	.	.
[Medication=2]	0.22	0.09	81.59	2.31	0.023	0.03	0.40
Wakeup_to_sim_hours	0.00b	0.00	.	.	.	.	.

a. Dependent Variable: Cortisol (transformed to Ln).
b. This parameter is set to zero because it is redundant

Pairwise Comparisons

(I) Simulation	(J) Simulation	Mean Difference (I-J)	Std. Error	df	p-value	95% Confidence Interval for Difference	
						Lower Bound	Upper Bound
None	Manikin	0.10*	0.04	132.13	0.014	0.02	0.18
	AR	0.06	0.04	134.67	0.122	-0.02	0.14
Manikin	None	-0.10*	0.04	132.13	0.014	-0.18	-0.02
	AR	-0.04	0.05	74.07	0.409	-0.13	0.05
AR	None	-0.06	0.04	134.67	0.122	-0.14	0.02
	Manikin	0.04	0.05	74.07	0.409	-0.05	0.13

Based on estimated marginal means

*. The mean difference is significant at the 0.05 level.

a. Dependent Variable: Cortisol (transformed to Ln).

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

II. Effect Modificators of AR simulation

II.a Cortisol:

Post-traumatic stress disorder (PCLC, p=0.389), baseline perceived stress (PSS_bl, 0.089), and baseline reported depression (CESD_bl, 0.514) failed to achieve statistical significance when introduced to the model predicting Salivary Cortisol based on AR.

Estimates of Fixed Effects

						95% Confidence Interval	
Parameter	**Estimate**	**SE**	**df**	**t**	**Sig**	**Lower Bound**	**Upper Bound**
Intercept	-1.68	0.28	108.64	-6.06	0.000	-2.23	-1.13
[Simulation=0]	0.06	0.04	75.33	1.45	0.153	-0.02	0.15
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[Day=1]	-0.06	0.06	73.94	-0.94	0.351	-0.18	0.06
[Day=2]	0.00b	0.00	.	.	.	.	.
[Sex=1]	0.19	0.10	78.36	1.89	0.063	-0.01	0.39
[Sex=2]	0.00b	0.00	.	.	.	.	.
[Medication=1]	-0.04	0.12	82.62	-0.37	0.711	-0.28	0.19
[Medication=2]	0.00b	0.00	.	.	.	.	.
Wakeup_to_sim_hours	-0.05	0.02	147.51	-1.96	0.051	-0.09	0.00
PCLC	0.01	0.01	75.71	0.87	0.389	-0.01	0.02
pss_bl	-0.02	0.01	81.50	-1.72	0.089	-0.05	0.00
CESD_bl	0.01	0.02	77.97	0.66	0.514	-0.02	0.04

a. Dependent Variable: Cortisol (transformed to Ln).
b. This parameter is set to zero because it is redundant.

II.b Galvanic Skin Response

Post-traumatic stress disorder (PCLC, p=0.079), baseline perceived stress (PSS_bl, 0.126), and baseline reported depression (CESD_bl, 0.257) failed to achieve statistical significance when introduced to the model predicting GSR based on AR.

Estimates of Fixed Effects

						95% Confidence Interval	
Parameter	**Estimate**	**SE**	**df**	**t**	**Sig**	**Lower Bound**	**Upper Bound**
Intercept	3.49	0.52	1421.73	6.70	0.000	2.47	4.51
[Simulation=0]	-0.30	0.06	192.22	-4.66	0.000	-0.42	-0.17
[Simulation=2]	0.00b	0.00	.	.	.	.	.
[Day=1]	-0.24	0.14	106.67	-1.74	0.084	-0.51	0.03
[Day=2]	0.00b	0.00	.	.	.	.	.
[Sex=1]	0.56	0.18	480.35	3.03	0.003	0.20	0.92
[Sex=2]	0.00b	0.00	.	.	.	.	.
[Medication=1]	0.26	0.22	688.79	1.15	0.249	-0.18	0.69
[Medication=2]	0.00b	0.00	.	.	.	.	.
Wakeup_to_sim_hours	-0.06	0.05	2651.24	-1.35	0.178	-0.15	0.03
PCLC	-0.03	0.01	512.01	-1.76	0.079	-0.06	0.00
pss_bl	0.04	0.02	583.18	1.53	0.126	-0.01	0.08
CESD_bl	0.03	0.03	500.48	1.14	0.257	-0.02	0.09

a. Dependent Variable: MeanSC_Long_SQRT.
b. This parameter is set to zero because it is redundant.

III. Discussion

Both the manikin and AR simulators elicited emotional (i.e., a reduction in positive emotion and an increase in negative emotion) and stress responses during and after the simulations. This is consistent with previous studies that showed that simulation in medical education can elicit a stress response as well as a range of emotional and cognitive changes. Given previous research showing stressors can enhance learning, this suggests both modalities of simulation can have a beneficial effect for learners, although future studies will need to evaluate actual learning outcomes. Of note, there was some concern that AR might be associated with a dangerously high level of stress because of the added realism and interactive nature, however, it does not seem to be any more stressful than past medical training approaches adding some indication that this is not a concern at least in this context. Further sub-analysis of participants with pre-existing PTSD, perceived stress and depression did not show statistically significant differences in stress with AR simulation, suggesting that even those with pre-existing conditions do not need to be excluded from AR technology. Further, the levels of stress and negative emotion reported in these simulations do not appear to be at levels that are different than other study averages.

Augmented reality technology is relatively new and its ability to elicit a stress response when compared to standard manikin simulation technology could help guide future educational practices and research. Our study shows no significant differences between AR and standard manikin simulation technology, except a small difference where the increase in skin conductance in response to the manikin simulator was significantly higher than that of AR. Cortisol differences, however, were not different across the two platforms. This suggests that the AR approach is psychologically and physically comparable to standard manikin-based simulators, and perhaps even less physiologically arousing than past learning modalities.

Given the nature of the simulations involving pediatric deaths it is not surprising that the overall stress increased during and after each simulation. However, students show decreased stress levels in their second simulation. Previous studies have shown that stress factors in simulation-based training may help with the acquisition of stress management skills. In addition to stress management skills, it could suggest a desensitization to the simulation regardless of type of simulator. Chang et al. suggested that VR simulation could be used to desensitize pediatric physicians from stressful situations based on his study evaluating VR stress response and real-life situations. Although, Hardernberg, et al. showed no decreased stress response in nursing students with repeated simulations. Our results contradict this, which could be very useful for educational efforts and future training for many types of medical practitioners who experience high stress situations.

III.a Limitations

This is a single site study comparing AR simulation to standard manikin-based simulation. While we attempted to look at multiple evaluators of emotional stress (cortisol, self-reported stress, electrodermal activity) these still may not have fully captured the stress response of the students. Furthermore, in an attempt to identify differences in stress response we intentionally used a very high stress case. Other simulation cases that are not as high-stakes likely elicit lower stress levels.

Furthermore, cortisol and electrodermal activity have confounding factors such as gender, medications, and time since waking. While corrected for these confounders there may be others unaccounted for that resulted in noise in the data.

III.b Conclusions & Implications

Augmented reality simulators elicited similar stress responses to manikin-based simulators suggesting they are comparable tools for medical education. Furthermore, there was no evidence of AR simulators causing excessive stress to participants. Future research should evaluate whether AR simulators increase learning outcomes or help with desensitization or stress management skills with repeated use.

III.c Manuscript

A Manuscript was prepared for submission to an Academic Journal. The Manuscript contains references to previous research that was accomplished.