Does gamified breath-biofeedback promote adherence, relaxation, and skill transfer in the wild?

Dennis R. da C. Silva, Ricardo Gutierrez-Osuna, *Senior Member*, IEEE

Abstract—This paper investigates whether gamification of deep breathing (DB) exercises promotes relaxation, skill transfer, and adherence to treatment in ambulatory settings. We designed a game-biofeedback (GBF) intervention where users perform DB exercises while playing a video game, and the game adapts according to the user's breathing rate using negative reinforcement instrumental conditioning. As a control, we developed an interactive paced-breathing treatment (PACE) where users follow a visual signal with their breathing and touch. In a user study, 30 participants were randomly assigned to GBF or PACE, and were allowed to practice at their leisure over the course of three days. Results show that the GBF group practiced their treatments significantly more often, achieved better skill transfer at post-test, and obtained a higher increase in self-reported positivity and relaxation during treatment. Our findings suggest that the use of negative reinforcement coupled with a fun casual game can be used as an alternative tool to promote relaxation and improve adherence to stress management interventions.

Keywords—Relaxation training, game biofeedback, deep breathing, skill transfer, wearable sensors, stress management.

1 Introduction

Deep breathing (DB) is an effective self-regulation strategy that has been used for millennia to achieve relaxation [1]. Lowering one's breathing rate increases cardio-respiratory synchronization, which prompts the mind and body to shift to a parasympathetic-dominant state [2]. In fact, under certain conditions, DB can produce nearly-complete inhibition of the sympathetic (i.e., fight-or-flight) system [3]. Despite its effectiveness and simplicity, DB has two major limitations. First, DB is commonly practiced in quiet settings, so the skill may not transfer to real-life situations under stress, as Stress Exposure Training suggests [4]. Further, DB interventions suffer from high dropout rates [5] because they lack engaging elements, and can become monotonous [6].

An alternative technique to improve self-regulation is biofeedback. In biofeedback, a patient wears a sensor that measures a target physiological signal (e.g., heart rate, breathing rate) while monitoring that signal in real time, typically in a visual display. This allows patients to become aware of physiological responses that otherwise would be unnoticed. Biofeedback offers several benefits. First, it allows users to become more aware of their body, an important aspect of self-regulation [7] known as interoceptive awareness [8]. Second, biofeedback allows users to observe how different behaviors affect their physiology, helping them develop self-regulation skills. Further, the wide availability of physiological sensors (e.g., fitness trackers), makes biofeedback an affordable treatment that can be practiced in the comfort and privacy of one's home [9]. However, biofeedback can become repetitive, and thus difficult to adhere to in the long run.

A potential approach to improve adherence is to *gamify* DB exercises through biofeedback [10]. Video games provide multiple mechanisms to increase engagement, such as storytelling, game mechanics, visual elements, and dynamic difficulty adjustment [11, 12]. As an example, DB exercises may be embedded into video games to (1) make the practice more engaging and (2) practice in the presence of a mild stressor [13, 14]. In a series of prior studies [14-16], we developed a game-biofeedback (GBF) technique where a video game adapts in real-time, rewarding players for maintaining a low breathing rate during gameplay. Under controlled laboratory conditions, we showed that GBF improves transfer of DB skills to stressful tasks.

Motivated by these findings, we recently conducted a pilot study to validate our game-biofeedback concept in an ambulatory setting [17]. We used a 2×2 factorial design with breathing guide (via respiratory biofeedback vs. a respiratory pacing signal) and gamification (game vs. no game) as independent factors. Participants were randomly assigned to one of four groups: (1) visual biofeedback, (2) paced breathing, (3) game biofeedback, and (4) game with paced breathing. Then, they were asked to practice DB as they saw fit for the next 6-8 hours, while carrying out their daily activities. We found that participants in the two game groups practiced with the treatment more often than those in the non-game groups (visual biofeedback, paced breathing), suggesting that gamification of DB exercises

- *D. R. da C. Silva is with the Department of Computer Science at Texas A&M University, College Station, TX 77843. E-mail: silva.dennis@tamu.edu.*
- *R. Gutierrez-Osuna is with the Department of Computer Science at Texas A&M University, College Station, TX 77843. E-mail: rgutier@tamu.edu.*

may be a useful strategy to increase adherence. Further, self-report measures indicated that the game-based interventions were more successful at promoting skill transfer (i.e., DB during stressful tasks), whereas the non-game interventions were better at promoting in-the-moment relaxation (i.e., DB during practice).

While these results are encouraging, the pilot study [17] had several limitations that the current study seeks to address. First, improvements in skill transfer were only significant for self-report measures, but not for breathing rates during the stressor. These results could have been due to dosage effects: the pilot study lasted 6-8 work hours in a single day, which did not give participants ample opportunities to practice. Second, measuring adherence (i.e., number of practice sessions) in such a short period is problematic, as it does not allow novelty effects to wear off. Finally, the pilot study measured (self-reported) feelings of in-the-moment relaxation at the end of the day, rather than immediately after each treatment session. As such, these measures may have suffered from memory recall issues, i.e., participants had to recall their overall feelings across multiple sessions completed earlier in the day.

This article presents results from a new three-day ambulatory study that seeks to address the limitations of our pilot study [17]. In this new study, participants were randomly assigned to one of two conditions (GBF vs DB) and were asked to self-report their emotional state (valence and arousal) immediately before and immediately after they completed a practice session. This new study design allows us to answer several related questions: Does adherence drop significantly over time? If so, to what extent is adherence due to engagement? Do objective measures of skill transfer (i.e., reduced breathing rate during a stressor) differ between GBF and DB if participants are allowed to practice more times? Are there differences in use patterns (e.g., times, locations, contexts) between the two interventions? Do the two interventions have different effects in terms of in-the-moment vs. sustained relaxation? To answer these questions, we tested the following hypotheses:

- ***H1**:* GBF has higher adherence to treatment than DB
- ***H2**:* Differences in adherence between GBF and DB are mediated by engagement
- ***H3**:* GBF outperforms DB in terms of skill transfer, both immediately after the laboratory treatment session (short-term skill transfer; *H3a*) and after the ambulatory treatment is completed (three-day skill transfer; *H3b*)
- ***H4**:* In contrast, DB outperforms GBF in terms of in-the-moment relaxation

2 Related work

2.1 Game biofeedback to *self-regulate* breathing

Multiple studies have shown that combining respiratory biofeedback with games can improve breathing [13, 18, 19]. Sonne and Jenson [18] proposed a game (ChillFish) that required children to maintain a low breathing rate (BR) in order to keep a puffer fish inflated so it can collect rewards. The authors found that heart rate variability (HRV)—a physiological marker of relaxation — was significantly higher when playing ChillFish than with other activities in the study (e.g., talking, playing Pacman). Using a similar strategy, Shih et al. [13] developed an adaptive game where the speed of a sailboat increased at low BRs. The game was rated as significantly more enjoyable than a control condition (paced breathing) and led to higher HRV. More recently, Schlatter et al. [20] had medical students perform one of three 5-min exercises (paced breathing, HRV-biofeedback, control) prior to performing a critical medical procedure in a simulator. Students in both experimental groups reported lower levels of stress than those in the control group and received higher ratings of task performance by assessors blinded to group allocations.

Prior studies have also used virtual reality (VR) to increase immersion and engagement in breathing exercises [21-23]. For example, Brammer et al. [21] developed a VR biofeedback game in which players (police officers) had to shoot hostile zombies and save benign ones, all the while breathing slowly to improve peripheral vision in the game. Using a within-subject design, the authors found that BRs were significantly lower during sessions played with biofeedback compared to sessions without biofeedback. Weerdmeester et al. [23] developed a VR biofeedback game (DEEP) in which players navigate through an underwater world whose surroundings mirror their breathing cycle. The authors conducted a user study to compare DEEP against paced breathing (control) as a treatment for anxiety. Both interventions significantly reduced self-reported anxiety (pre- vs. post-test), and the decrease was stable at a 3-month follow up. However, differences in anxiety between groups at post-test and follow-up were not statistically significant, indicating that DEEP was not superior to control. Engagement ratings were significantly higher for the DEEP group on the first session but reached similar levels to the control group as sessions progressed. This suggests that novelty effects played a significant role.

2.2 Game biofeedback for *skill transfer*

Research on gamified biofeedback for skill transfer dates to the 1990s. In early work, Larkin et al. [24] studied whether heart-rate (HR) biofeedback could lead to skill transfer. The authors used a 2×2 factorial design, with biofeedback (yes/no) and scoring[1] (based on game performance alone vs. game performance *combined with* low HR) as independent factors. Compared to players whose score was based on game performance alone, players who received the combined score had significantly lower HRs when playing the game *without* biofeedback and completing a novel mental arithmetic task.

More recently, our group developed Chill-Out [14], a biofeedback game that applied a penalty (increased challenge) when participants' BR exceeded 6 breaths/min. To measure skill transfer, we compared reductions in BRs when participants completed a stressful task (Stroop

[1] The final score was based either on (1) participants' performance on the game, or (2) game performance and participants' ability to maintain a low heart rate.

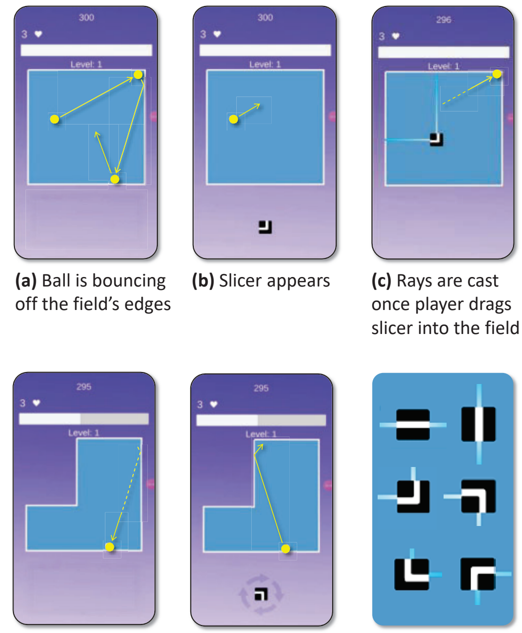
Color-Word Test (referred to as Stroop in this manuscript) [25, 26] before and after training. The biofeedback game led to a higher reduction in BR (pre-post treatment) than a control condition. In a follow-up study [15], we compared skill transfer and skill acquisition speed (number of training sessions needed to acquire DB skills) when biofeedback was provided (1) in a visual form, (2) by adapting the game, and (3) as a combination of the two. We found that game adaptation outperformed visual feedback in skill acquisition and transfer, and that the combined strategy outperformed either form of biofeedback in isolation. In a final study [16], we examined whether partial reinforcement (e.g., applying penalty 75% of the times) would improve resistance to extinction compared to continuous reinforcement (e.g., applying the penalty 100% of the times), and whether this would come at the cost of additional training sessions. Partial reinforcement increased resistance to extinction, measured as how long BRs and electro-dermal activity (EDA) remained low after removing the biofeedback signal. Surprisingly, partial reinforcement did not require more training sessions than continuous reinforcement.

2.3 Game biofeedback in ambulatory settings

Aside from our pilot study [17], a handful of studies have examined gamified biofeedback into the wild [17, 27, 28]. Yahav and Cohen [27] conducted an 8-week ambulatory study examining whether EDA-based biofeedback games could alleviate anxiety and behavior symptoms (e.g., sleeping problems, fits of crying) in adolescents. Participants practiced relaxation with three biofeedback games, where characters in the game advanced once EDA decreased to a certain level. Ratings of state anxiety, test anxiety, behavior symptoms, and self-esteem improved significantly for the treatment group compared to a control group (which did not receive biofeedback training). Osman et al. [28] developed Botanical Nerves, an HRV biofeedback game to track stress. The game presents a tree that becomes greener and with more abundant leaves as the player's HRV increases and becomes wilted as HRV decreases. In a 10-day study (5 days without biofeedback, and 5 days with biofeedback, counterbalanced across participants), the authors found that participants' trees were healthier and greener on biofeedback days, indicating that biofeedback promoted reductions in stress.

3 METHODS

We describe the current study by first introducing the base game that served as the platform for gamification. Then, we describe the two interventions in the study: an experimental treatment (gamified breath biofeedback), and the control treatment (paced breathing). Next, we describe the tasks we used to elicit stress and measure skill transfer. This section concludes with a detailed description of the experimental protocol.



(a) Ball is bouncing off the field's edges **(b)** Slicer appears **(c)** Rays are cast once player drags slicer into the field

(d) Intersecting region is removed **(e)** A new slicer appears, and the cycle repeats **(f)** Different types of sliders can appear

Fig. 1. **Scale base game: (a)** The UI shows a field with a ball bouncing from the edges. **(b)** A slicer appears underneath the field for players to drag into the field. **(c)** Once the slicer is placed, it casts rays toward the edges of the field. **(d)** The portion of the field that does not contain the ball is removed. **(e)** A new slicer appears as the ball continues to bounce from the edges of the modified field. **(f)** Different types of sliders are available in the game.

3.1 Base game

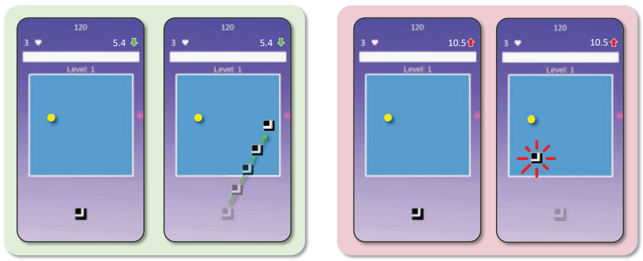
Our base game is a clone of *Scale*, an "endless" mobile game that is available on multiple platforms[2]. Illustrated in **Fig. 1a**, the game presents a bouncing ball on a square field, and the objective is to progressively reduce the size of the field. For this purpose, the player drags a *slicer* onto the field –see bottom of the screen in **Fig. 1b**. Once the slicer is released, rays from its ends are cast until they intersect the field's boundaries (**Fig. 1c**), and the portion that does *not* contain the ball is removed (**Fig. 1c**). One of several types of slicers (see **Fig. 1f)** is randomly loaded whenever the player completes a move.

To advance to a new level, the player must reduce the field to at least 50% of its original size. At that point, the new level starts with the final shape of the previous level, enlarged to match the width or height of the screen. Players start at level 1 with three lives and receive three additional lives every time they advance to a new level. However, players must place the slicer carefully: if the ball collides with the rays being cast, the player loses one life. The game is considered "endless" in that the player can advance to new levels indefinitely, until they lose all their lives. At that point, the game restarts at level 1 again.

3.2 Treatments

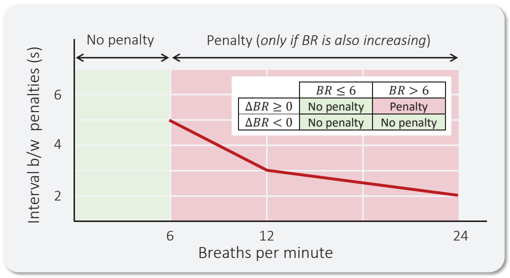
The two treatments in the study share a common goal: encouraging participants to breathe at 6 breaths/min (0.1Hz), a resonance frequency of the cardiorespiratory system that has been shown to maximize HRV [17, 29].

[2] Though the game is available on the Google Play store, we developed a custom version in Unity to be able for the game to adapt based on the player's breathing game, as we describe in the next section.



(a) When breathing slowly (BR<6), the user is allowed to drag the slicer

(b) When breathing fast (BR>6), the game drags the slicer randomly



(c) Interval between random drags as a function of breathing rate

Fig. 2. **Treatment 1: GBF:** The player's BR is indicated on the top-right corner, along with an arrow indicating whether BR is increasing or decreasing. **(a)** When the player's BR<6 and/or decreasing, the biofeedback game gives the player full control of the slicer. **(b)** When the player's BR>6 and increasing, the biofeedback game places the slicer at a random location (negative reinforcement). **(c)** Time between events (slicers placed randomly onto the field) as a function of BR.

3.2.1 *Treatment 1: Game biofeedback (GBF)*

The first treatment is an adaptive version of *Scale* that trains participants to breathe slowly through negative reinforcement instrumental conditioning [30]. In negative reinforcement, a target behavior is reinforced by removing negative consequences, e.g., a driver fastens the seat belt [behavior] to stop the car's beeping noise when driving [negative consequence]. In our case, participants must breathe slowly (i.e., behavior to be strengthened) to prevent the game from behaving erratically (i.e., negative consequences).

The game adapts in response to the player's breathing rate, as illustrated in **Fig. 2c**:

- If BR is below target ($BR \leq 6$) or decreasing ($\Delta BR < 0$), the game defaults to the base game in **Section 3.1**, i.e., there are no penalties; see **Fig. 2a**.
- If BR is above target ($BR > 6$) and increasing ($\Delta BR \geq 0$), the game penalizes players by placing slicers automatically at random locations; see **Fig. 2b**.

We combined our prior experience with biofeedback games [14-17, 29, 31] with the following rationale to design the adaptation mechanism:

- Penalties are applied sharply at the $BR > 6$ boundary to alert the player, but the interval is large enough (5 seconds) to keep the game relatively easy to play.
- As BRs increase, the interval becomes shorter at a rate of 1/3 sec for every increase of 1 breath/min, until BRs reach 12 breaths/min, which is considered the lower bound of normal breathing (12-20 breaths/min) [32]. At a $BR = 12$ (3-sec interval), the game is hard to play.
- For BRs between 12-24 breaths/min, intervals decrease with a slower slope of 1/12 sec for each increase of 1 breath/min. The upper limit of 24 breaths/min provides a comfortable buffer above the upper limit of normal BRs (20 breaths/min).
- At any point in the game, penalties are suspended if the player's breathing rate is decreasing $\Delta BR < 0$). This provision is key, as it ensures players have an opportunity to recover and reach the 6 breaths/min target at their own pace.
- Players must complete each level within 15 seconds, otherwise they lose one life. Without this time limit, players can stop placing slicers until they bring their BR below threshold. This defeats the purpose of the intervention, which requires players to control their BR under stress.

(a) The user follows a horizontal pacing bar with their breath and with touch interaction

(b) The bar changes color it the user stops touching it

Fig. 3. **Treatment 2: PACE:** The app provides a pacing signal in the form of a horizontal bar moving up and down the screen. **(a)** As long as the user follows the pacing signal with their finger, the app continues to count down time. **(b)** If the user stops following the pacing bar with their finger, the bar changes color, the app starts vibrating and the countdown timer stops.

3.2.2 *Treatment 2: Paced Breathing (PACE)*

As a control condition, we developed a mobile app for paced breathing (PACE). Illustrated in **Fig. 3**, PACE asks participants to synchronize their breathing by inhaling as a horizontal bar moves up, and exhaling as it moves back down. To ensure a fair comparison against the GBF app, the PACE app also requires participants to pay attention to and interact with the screen, and in a way that encourages them to practice deep breathing, as follows.

Prior studies have shown that breathing can be *entrained* by other motor activities [33], such as walking [34], running [35], or even finger flexion and extension [36]. To take advantage of this phenomenon, PACE requires participants to follow the horizontal bar by touching it with a finger, as illustrated in **Fig. 3a-b**. If they do not (i.e., the distance between their finger and the bar exceeds 10 pixels), the bar's color changes from green to red, and the phone vibrates for 100ms every 5 seconds or until the bar starts being tracked again; see **Fig. 3c**. By combining touch interaction and the entrainment mechanism, PACE requires similar levels of attention and interaction as GBF. *We deemed this to be critical, to ensure that any between-group differences in outcome measures (adherence, engagement skill transfer, and in-the-moment relaxation) would only be due to practicing deep breathing either (1) with biofeedback and a mild challenge (i.e., the game), or (2) without both.*

3.3 Stressors

Following prior work [15-17], we used two additional tasks as stressors: the Stroop color word test and mental arithmetic. Participants completed both tasks before and after treatment, and we used differences in BRs as an objective indicator of skill transfer.

Illustrated in **Fig. 4a-b**, the conventional (incongruent) Stroop test asks participants to press one of several buttons at the bottom of the screen based on the *font color* of a word displayed at the top, ignoring the *text* of the word. This requires participants to inhibit the prepotent response of reading the word, leading to stress. To reduce learning effects, we implemented a custom version of the Stroop that introduced additional challenges. First, the app changes arbitrarily between asking players to choose based on color (**Fig. 4a**) or based on text (**Fig. 4b**). This forces players to readjust their strategy constantly. In addition, the position of the buttons at the bottom changes arbitrarily each round. Finally, the app also startles the player by sounding a loud buzzer if they select the wrong choice, or time runs out for that round.

The second stressor is a mental arithmetic task that increases cognitive load [37, 38]. Following prior work [15-17] we used the King of Math app available on Google Play [39], as it offers a variety of arithmetic tasks and difficulty levels. Participants play the game in the *mixed* setting, which includes arithmetic (**Fig. 4c**) and statistical questions (**Fig. 4d**). Participants start in chapter 1 of the mixed setting and progress through the chapters until they complete them, or time runs out. The difficulty level increases as participants clear chapters, making this task adapt to each participant's skills. Unlike Stroop, however, there is no time pressure to complete each arithmetic task. Instead, we ask participants to achieve a score as high as possible.

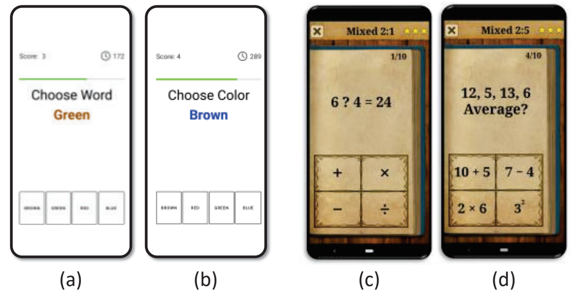


Fig. 4. **Stressors in our study**. Stroop color word test with two different instructions: **(a)** choose word, and **(b)** choose color. King of Math: examples of **(c)** arithmetic, and **(d)** statistical questions.

3.4 Experimental protocol

To test our four hypotheses (see Introduction), we designed an experimental protocol consisting of three phases: a pre-test phase in the lab (morning of day 1), an ambulatory phase (afternoon of day 1, all day 2, morning of day 3), and a post-test in the lab (afternoon of day 3) – see **Fig. 5**.

3.4.1 Pre-test phase

Upon arrival to the lab on day 1, participants provided informed consent, and were fitted with two sensors:

- A Zephyr BioHarness 3.0 chest strap [40] that measures heart rate and respiration rate, and streams data via Bluetooth. Participants wore the sensor for the duration of the study (10-12 hours/day).
- A Biograph Infinity [41] system that measures EDA[3], a selective marker of arousal [43]. Unlike the BioHarness, however, the EDA sensor is only worn in the lab (i.e., pre-test and post-test phases).

Once sensors are attached, participants filled out a questionnaire (Pre-Q) about demographics, experience with relaxation techniques, state-trait anxiety inventory (STAI) [44], and a personality inventory (short Big-Five) [45]; see Supplementary Materials, Appendix 7. Then, participants completed a series of tasks:

- **Pre-practice stressor (Stroop1):** Participants perform the Stroop task for 5 min, with a 3-sec response time per prompt. This task provides a measure of participants' stress reactivity prior to treatment.
- **Paced Breathing (PB)**: Following prior work [17, 29], participants follow a pacing signal consisting of a 4-sec inhale and a 6-sec exhale[4]. This helps them get used to breathing at the target BR (6 breaths/min).
- **Base game (Game)**: Participants play *Scale* for 5 min without biofeedback. This helps them get familiarized with the game mechanics and provides a baseline for the GBF group.
- **Laboratory Treatment (LT):** Participants practice with their treatment until they spend a *cumulative* 2 mins at the target BR (not necessarily consecutive) or 5 min had elapsed[5]. If the 5-min timer runs out, participants are asked to repeat the session until they succeed. This ensures participants can perform their treatment in the lab, before attempting it in the wild.
- **Post-practice stressor (Stroop2):** Participants perform the Stroop task for 3 min, with a shorter response time of 2 sec, to reduce potential learning effects from Stroop1. This task provides a short-term measure of skill transfer, by comparing BR to those during Stroop1.
- **Post-practice stressor (Math1):** Participants perform mental arithmetic for 3 min. This task allows us to measure short-term skill transfer to a new stressor.

3.4.2 Ambulatory phase

After completing the lab session and removing the EDA sensor, participants are dismissed for the next 3 days, with the suggestion they complete multiple treatment sessions,

[3] We used pre-gelled Ag/AgCL electrodes attached to the sole of participants' right foot, instead of the hand, which would cause motion artifacts when participants complete tasks on a smartphone. A prior study had demonstrated the effectiveness of measuring foot EDA as an alternative to palms or fingers [42] W. Boucsein, *Electrodermal activity*, 2 ed. New York, NY: Springer, 2012..

[4] We used an exhalation/inhalation ratio greater than one (6/4), as prior work shows it leads to higher relaxation states [46] G. Strauss-Blasche, M. Moser, M. Voica, D. McLeod, N. Klammer, and W. Marktl, "Relative timing of inspiration and expiration affects respiratory sinus arrhythmia," *Clin Exp Pharmacol,* Aug. 2000..

[5] The app has a 2-minute countdown timer that only decreases when the participant's breathing rate is at or below target, and the timer must reach zero in less than 5 minutes.

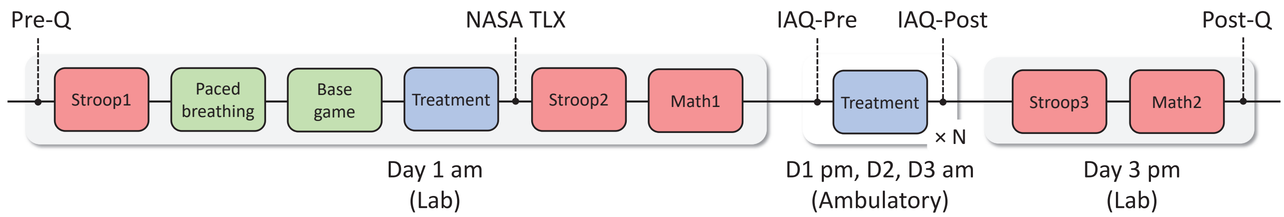


Fig. 5: **Overview of the experimental protocol**. Red boxes are stressors; green boxes are training steps; blue boxes are treatment sessions. In the morning of day 1 (D1), participants fill out a Pre-Questionnaire (Pre-Q), then complete Stroop1, a Paced Breathing session, practice with the Base Game, perform their respective Treatment in the lab, and complete two stressors: Stroop2, and Math1. For the next 2.5 days, participants complete as many ambulatory Treatments as they wish. Prior to and after each ambulatory Treatment session, they complete in-app questionnaires (IAQ-Pre, IAQ-Post). In the afternoon of D3, participants return to the lab to complete Stroop3, Math2, and a Post-Questionnaire (Post-Q).

whenever it is a good time to take a short break, or they feel stressed or agitated. *We refrain from asking participants to complete a fixed number of sessions, since this is one of the dependent variables of the study*. On the evenings of days 1 and 2, participants receive email reminders to charge the phone and BioHarness overnight.

Treatment sessions completed ambulatorily are identical to the sessions completed in the lab phase: for a session to be considered valid, participants must spend a cumulative 2 mins at the target BR during the 5-min session duration, otherwise the session is considered invalid and is not considered in the analysis. Immediately before and immediately after completing each ambulatory session, participants complete an in-app questionnaire (IAQ) where they rate their valence and arousal pre- and post-treatment (5-point scale) as well as their experience with the game, in terms of engagement, ease of reaching the target breathing rate and ease of remaining focused on the app (see Supplementary Materials, Appendix 8). These questions allow us to evaluate *in-the-moment* changes in affect after each session and capture participants' experience throughout the study.

3.4.3 Post-test phase

On the evening of day 3, participants return to the lab, where they are fitted with the EDA sensor and asked to complete two tasks:

- **Stroop3:** Participants performed the Stroop for 3 min, with a response time of 2 sec (as in Stroop2).
- **Math2:** Participants perform the math task for 3 min.

After completing these two tasks, participants fill out a post-study questionnaire (Post-Q) about their experience during the study, thoughts about their treatment, and suggestions for future experiments (see Supplementary Materials, Appendix 9).

3.5 Study enrollment

We enrolled 30 healthy participants (15/15 f/m; 25±4.5 years) via Texas A&M University (TAMU) bulk mail service. We randomly assigned 15 participants to each group (GBF: 8/7 f/m; PACE: 7/8 f/m). All participants enrolled in the study were compensated $75 for their participation. All participants were TAMU students and were enrolled based on the following inclusion criteria: young adult (18-35 years of age), with no self-reported history of anxiety or depression, and fluent English speaker. The study was approved by the TAMU Institutional Review Board (#IRB2019-0218D). All participants provided written consent before taking part in the study. Study questionnaires are included as Supplementary Materials, Appendices 7-9.

To determine the sample size used in the present study, we performed a power analysis based on physiology data from an earlier study that compared different types of biofeedback for relaxation training [47]. In that study, the average BR (± standard deviation) at pretest was 18.07 ± 3.97 for GBF and 17.51 ± 7.43 for Control, whereas their respective averages at post-test were 5.46 ± 1.55 (GBF) and 17.52 ± 3.05 (Control). The first step to estimate the sample size was to obtain the effect size using the Cohen's method [48] with the data reported in the prior study. Specifically, we 1) computed the differences in BR between post-test and pretest for each group, 2) calculated the mean difference in BR for each group, 3) computed the pooled standard deviation of the BR differences, and finally, 4) applied Cohen's method, resulting on an effect size of 1.2. We then used the G*Power software [49] to estimate the required sample size, using power of 0.8, alpha of 0.05, and the estimated effect size of 1.2, resulting in a sample size of 12 participants per group. A similar analysis was done in our prior work [50]. We decided to enroll 15 participants per group to have a bit of a buffer.

4 Results

4.1 Participant sample

In a first step, we examine participants' responses to the pre-study questionnaire (Pre-Q), as well as the anxiety and personality instruments.

Many participants reported some prior experience with mindfulness or meditation (18/30), but very few practiced it regularly (2/30). Likewise, many participants had prior experience with DB (20/30), but none practiced it regularly. Finally, the vast majority had no experience with biofeedback (28/30), and none practiced it regularly.

The STAI scores (see Supplementary Materials, Appendix 1.1) indicate that most participants had either "low to no anxiety" or "moderate anxiety", and none had "high anxiety". This result is consistent with our recruitment criteria, which excluded participants with a self-reported history of anxiety. Further analysis showed no significant differences between the two groups for state or trait anxiety. Finally, scores on the short Big-Five Inventory (see Supplementary Materials, Appendix 1.2) showed no significant differences between groups for any of the dimensions.

Together, these results indicate that participants in both groups were similar in terms of (un)familiarity with relaxation techniques, personality traits, and anxiety.

4.2 Validity of the protocol

To validate our experimental protocol, we examined participants' physiological responses and self-report measures to the various tasks. Results for the protocol validation analysis are shown in **Fig. 6** and Supplementary Materials, Appendix 3.1.1

Breathing rate. BRs for both groups across tasks are shown in **Fig. 6** (top panel). During Stroop1, BRs were within 17-23 breaths/min, slightly higher than normal (12-20 breaths/min [51]), as expected since the Stroop task is a mild stressor. Both groups had similar BRs, which further confirms they came from the same population. During Pace Breathing, both groups reached and maintained the target BR, indicating the task was effective as a training tool. During the Base Game training session, BRs returned to similar levels as Stroop1, indicating that playing the base game elicited stress, also as intended. During the Laboratory Treatment session, both groups reached the target BRs, indicating both interventions promote DB, also as intended[6]. Finally, during the Ambulatory Treatments, BRs for both groups were similar during the lab and ambulatory treatments.

Electrodermal activity. EDA provides an independent measure of stress that is not directly manipulated during the interventions. As such, it can be used to corroborate the results obtained from analyzing BRs. To quantify EDA, we computed the number of skin conductance responses (#SCRs) per minute, which are associated with sudden activation of the sympathetic nervous system [52]. EDA results are also shown in **Fig. 6** (bottom panel) and Supplementary Materials, Appendix 3.2. As expected, we do not observe differences in EDA between groups in the tasks that precede the Laboratory Treatment, further corroborating that both groups were from the same population. Further, EDA is consistent with the nature of those three tasks: (1) it is highest during Stroop1, indicating the task leads to a stress response; (2) drops to its lowest value during DB, indicating a relaxation response, and (3) returns to a high during Base Game, indicating the game also leads to a stress response. Finally, both Laboratory Treatments reduce EDA to similar levels as those during PB for both groups.

Task load. During the pre-test phase, participants completed the NASA Task Load Index (NASA TLX) after completing their respective Laboratory Treatments. Results in **Fig. 7a** show participants rated GBF to be more demanding than PACE, in terms of mental and temporal demand, effort, and frustration. Interestingly, participants in GBF rated their performance in the task higher than participants in PACE. Results in detail are shown in the Supplementary Materials, Appendix 1.3.

4.3 Outcome measures

Having validated the participant pool and experimental design, we examine the study's three outcome measures:

- **Adherence to treatment**, measured as the number of sessions practiced during the ambulatory phase;
- **In-the-moment relaxation**, measured by self-reported valence and arousal before and after each ambulatory session (primary) and BR/EDA while practicing the intervention (secondary); and
- **Skill transfer**, measured by BR/EDA during the stress tasks after the intervention.

To interpret these outcomes, we also examine engagement, as it is the assumed mediator for higher adherence, as well as qualitative feedback during the exit interviews.

4.3.1 Outcome 1: Adherence to treatment

Analysis of the app logs shows that GBF participants completed significantly more sessions (13.2 ± 4.0) than PACE participants (9.0 ± 3.2) across the three study days[7].

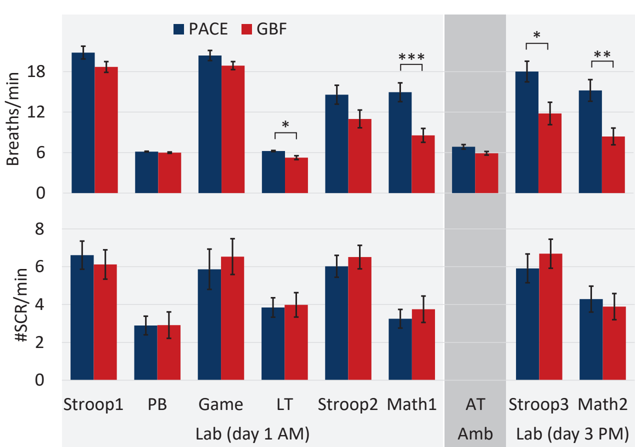


Fig. 6: Breathing rates and electrodermal activity for PACE and GBF during each study task. Electrodermal activity is only measured in the lab sessions, but not ambulatorily.

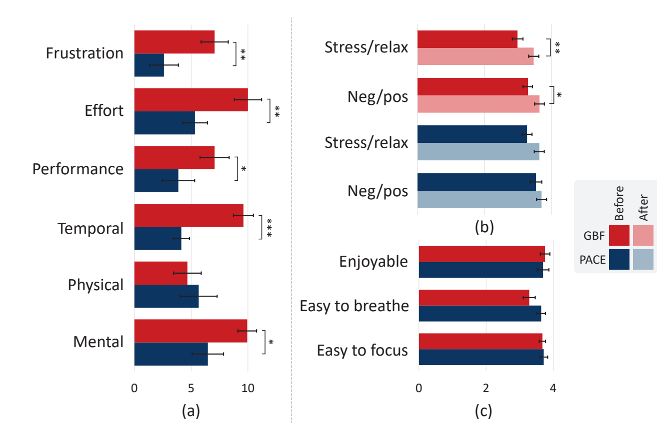


Fig. 7: **(a)** NASA-TLX results. The questionnaire was completed after the practice sessions for both treatments. In-app questionnaire (IAQ) ratings for **(b)** affect and **(c)** experience.

[6] We had expected that mastering GBF would require more training sessions than PACE (as GBF is a harder task), but differences between both groups were not significant. Thus, this result suggests that GBF is easy to learn.

[7] We used two-sample t-tests in all statistical significance tests, unless otherwise noted.

This result validates hypothesis *H1: GBF has higher adherence to treatment than DB*.

To determine if novelty effects could have influenced adherence, we also ran a 2-way ANOVA with intervention (GBF vs. PACE) and time (days 1, 2, and 3) as independent factors. We found a significant effect for intervention (more practice sessions for GBF), but not for time or interaction effects. Since the number of sessions for both groups did not change as the study progressed, we can rule out novelty effects. Results in detail are shown in the Supplementary Materials, Appendices 2.1 and 2.2.

4.3.2 Outcome 2: In-the-moment relaxation

Results for the questionnaire (IAQ) participants completed during each practice session are shown in **Fig. 7b-c**. The first three questions asked participants to assess valence and arousal before and after each session, following Russell's Circumplex model of affect [53]. In particular, the valence dimension allows us to determine if the interventions led to a depressed state (low arousal, negative valence) or to a calm state (low arousal, positive valence.) For each participant, we calculated the average rating of each question across all practiced sessions, then compared the ratings of each group via one-sample t-tests. GBF led to a significant increase in positivity and relaxation, whereas PACE only showed trends – See Supplementary Materials, Appendix 2.3. These results indicate that GBF is more effective than PACE in promoting in-the-moment relaxation, which leads us to reject hypothesis *H4 (DB outperforms GBF in terms of in-the-moment relaxation)*.

4.3.3 Outcome 3: Skill transfer

To assess skill transfer and test hypothesis H3 (GBF leads to better skill transfer than PACE), we examine both physiological measures (BR, EDA) separately.

Breathing rates. First, we measured short-term skill transfer (***H3a***) by comparing BRs under stress before treatment (Stroop1) vs. after treatment (Stroop2/Math1). BRs dropped significantly pre-post treatment for both groups, indicating both treatments promoted short-term skill transfer. Further, GBF had significantly lower BR than PACE during Math1, but not during CTW2. *Thus,* these results support hypothesis *H3a partially*, as GBF was superior to PACE only during Math1. Then, we measured three-day skill transfer (***H3b***) by comparing BRs under stress at the start of day 1 (Stroop1) vs. at the end of day 3 (CTW3/Math2). GBF reduced BRs significantly from Stroop1 to both Stroop3 and Math2. In contrast, PACE only reduced BRs from Stroop1 to Math2 (but not from Stroop1 to Stroop3). Further, GBF led to lower BRs than PACE in both tasks. These results support hypothesis *H3b*, indicating that GBF leads to better skill transfer than PACE when ambulatory treatment is conducted over a span of three days. Results in detail are shown in the Supplementary Materials, Appendices 3.1.2 and 3.1.3.

Electrodermal activity. EDA leads to conflicting interpretations when used as an indicator of skill transfer. On the one hand, differences between Stroop1 vs. Stroop2 (short-term), or between Stroop1 vs. Stroop3 (ambulatory) are not significant for either group, which suggests that neither intervention led to skill transfer. On the other hand, EDA during the Math1/2 stressors is significantly lower than during Stroop2/3 (also stressors) and comparable to that during Paced Breathing or the Lab Treatments, which induce relaxation. Why the two interventions lead to skill transfer for the mental arithmetic task but not to the color word test could be attributed to differences in demands between the two stressors. Although both tasks are mentally demanding, the Stroop task has the added element of significant time pressure: participants must respond within 2 sec, and maintain this pace for 3 min. Further, if participants do not respond within 2 sec or select the wrong answer, the app plays a loud buzzer sound, which is likely to elicit a startle response and the associated conditioned SCR. Results in detail can be found in the Supplementary Materials, Appendix 3.2.

4.3.4 Engagement

The last three IAQ items focused on engagement, asking participants to rate their experience after each session in terms of (1) enjoyment, (2) ease of reaching the target BR, and (3) remaining focused on the intervention. Results are shown in **Fig. 7c** and Supplementary Materials, Appendix 2.4. We had predicted that playing a casual game would be more ***enjoyable*** than tracking a bar on a screen. Surprisingly, we found no significant differences in enjoyment between the two interventions. We had also predicted that providing a pacing signal would make it ***easier to reach the target BR*** than playing the biofeedback game. Surprisingly as well, we found no statistical differences in terms of ease of reaching the target BR. Finally, we found no significant differences between the two interventions in terms of ***ease of remaining in focus***. This result is more nuanced. On the one hand, we had expected GBF would make it easier for participants to focus on the intervention (i.e., avoid distracting stimuli or thoughts). On the other hand, we also expected that PACE would make it easier for participants to focus on their breathing. Whichever interpretation of "focus" participants took when answering this question (refer to Section 4.3.5 for a qualitative analysis of participant feedback), these results suggest that adding a simple interactive element to DB (tracking the visual pacing signal with one's finger) is as effective in helping participants focus on the task as embedding DB exercises into a game.

Taken at face value, these results would lead us to reject hypothesis (H2), which states that differences in adherence between GBF and DB would be mediated by engagement. However, this coarse analysis does not consider the possibility that engagement ratings may have changed over the course of the study, which was also one of our research questions. To answer this question, we ran a 2-way ANOVA with intervention (PACE, GBF) and time (first day, last day) as independent factors. For ***enjoyment***, we found a significant effect for time (enjoyment decreased over time) but no significant effects for intervention or interactions. For ***ease of remaining focused***, we also found a significant effect for time (over time, it became harder to remain focused on the app), and no effect of intervention or interaction. For ***ease reaching the target BR***, we found no effects for intervention or time, but a significant crossover interaction (reaching the target BR became harder for PACE, but not for GBF). To corroborate these results, post hoc analyses between the first and last sessions indicate a significant decrease in the three measures for PACE, but not for GBF. *When taken as a whole, we find evidence both in support of and against hypothesis H2 (Differences in adherence between GBF and DB are mediated by engagement), indicating that our experiment fails to reject it.*

4.3.5 Qualitative feedback

Upon return to the lab on day 3, participants completed a post-study survey (Post-Q) that included (i) 5 quantitative and (ii) 5 qualitative questions; see Supplementary Materials, Appendix 9. The first set of questions asked participants to agree/disagree on a 5-point scale (1: completely disagree; 5: completely agree) that (1) reaching the target BR became easier over time, (2) focusing on the app became easier over time, (3) the app became more enjoyable over time, (4) find time to practice was easy, and (5) they looked forward to it. We found no statistical differences between the two interventions – see Supplementary Materials, Appendix 4. Participants agreed (ratings: 3.9-4.2) that their breathing, focus, and enjoyment improved over time, and to a lesser extent (ratings: 3.2-3.5) that they looked forward to and found time to practice.

Participant feedback for the second set of (open-ended) questions was more informative. The first question asked participants to describe the places, times, and situations in which they had practiced with the interventions. The most common **places** for practice were at home (50%) and at work/school (42%), with a strong preference for PACE participants to practice at home (57%) than at work (36%), and a smaller preference for GBF participants to practice at work (50%) than at home (42%). This suggests that participants sought to practice in a place that matched the characteristics of their intervention: quiet environments for PACE, and louder environments for GBF. When describing **times**, roughly half of the comments (46%) mentioned practicing at times of convenience (e.g., between tasks or classes, in their free time), with the remaining comments (54%) mentioning specific periods in the day (e.g., morning, evening). When comparing the two groups, there was a slight preference for GBF participants to practice at the start of the day, and for PACE participants to practice towards the end of the day. Finally, when describing **situations**, most comments (67%) mentioned practicing when stressed or anxious, and significantly fewer comments (29%) mentioned practicing when relaxed. Interestingly, PACE participants indicated a strong preference for practicing when stressed (83% of the comments), whereas GBF participants were more evenly split (50% when stressed, 42% when relaxed). In other words, participants' perceptions of the two applications were consistent with their intended purpose: PACE being primarily a relaxation tool, and GBF being a more versatile tool that intends to provide both relaxation and entertainment.

The second question asked participants to describe how using their respective app throughout the day had improved or worsened their day. We can draw several conclusions from participants' feedback. First, comments indicating the apps had improved the participants' day (79%) were far more frequent than those indicating the app had made the day worse (21%). The most frequent **positive** impact of the apps was in helping participants cope with stress or anxiety (38%), which occurred slightly more frequently for GBF (44%) than for PACE (33%). Other positive impacts included helping participants gather their thoughts (14%) and providing a diversion (8%). In contrast, the most frequent **negative** impact of the apps was by inducing stress or frustration (60%), which was noted more frequently in PACE (75%) than in GBF (50%). This result contradicts those from NASA TLX, which indicated that the GBF app was significantly more demanding than the PACE app. Thus, it appears that following (visually and by touch) a pacing signal was perceived as more stressful than playing a game designed to induce stress.

The third question asked participants the things they liked the most about the apps. We identified three major categories of comments, related to the treatment itself, physiological aspects, and mental aspects. Regarding the **treatment**, participants mentioned that it was easy (25%) – more frequently in PACE (33%) than in GBF (14%), and enjoyable (19%) –all in GBF (43%). These results are consistent with the app designs, since PACE has a simpler UI than GBF, and GBF is a videogame. Regarding **physiological** aspects, comments centered on the interventions helping participants focus on their breathing (36%) –PACE (43%) vs. GBF (29%), and gain control of their breathing (36%) –PACE (29%), GBF (43%). This indicates both apps helped participants become more aware of and gain control of their physiology, which relates to interoceptive awareness, an important component in emotion self-regulation [7]. Regarding **mental** aspects, the most common response was that the apps helped participants cope with stress (47%) –PACE: 50%; GBF: 45%.

The fourth question asked participants the things they liked the least about the app. We identified three categories for this question, regarding technical aspects, the intervention itself, or its effects. Regarding **technical** aspects, 64% of the comments mentioned sensor delay in estimating BR; most of them in PACE (7 comments) compared to GBF (2 comments). GBF participants were likely less bothered by sensor delays because their BR was displayed in the user interface while completing their sessions. Regarding the **intervention** itself, negative aspects were scattered (e.g., not very engaging, too difficult to achieve target BR), but several GBF participants (3 comments) mentioned the negative reinforcement (e.g., slicers being placed automatically, game becoming harder when breathing rate was increasing). Regarding the **effects** of the intervention, the most common set of negative comments referred to the app causing frustration or stress (56%) –PACE: 1 comment; GBF: 4 comments. This difference is aligned with NASA TLX but contrasts with the reports to the second question. As such, because of the conflicting evidence, it can be difficult to establish which group was under higher perceived mental stress when analyzing the open-ended responses in isolation.

The fifth question asked participants to suggest improvements to the apps. Participants made a wide variety of suggestions, but one category of comments stood out: **increase variety:** in-app features ("*more details about the breathing rate and other things can be added*"), visual aspects ("*could change the color of the board after every level*"), or in apps (*"can be introduced two apps and can find which type of task helps more?"*).

5 Discussion

We have conducted an ambulatory study to compare two mobile micro-interventions for stress self-regulation: gamified respiratory biofeedback (GBF) and paced breathing (PACE). For this purpose, we set to test four hypotheses, two related to adherence to treatment (H1, H2), and two related to the purported effects: skill transfer (H3) and in-the-moment relaxation (H4).

5.1 Adherence to treatment (H1, H2)

Our first hypothesis (H1) was that GBF would have higher adherence to treatment than PACE. GBF participants completed significantly more sessions than PACE participants, which confirms this hypothesis. This is an important finding, as it provides a potential solution to the high attrition rates of traditional relaxation and stress management interventions [6, 13]. Our findings are consistent with prior studies [17, 54, 55] that found increased adherence for gamified healthcare interventions.

Our second hypothesis (H2) was that differences in adherence would be mediated by engagement. Participants rated both interventions as equally engaging, which initially led us to reject H2. This is an unexpected result, since GBF delivers deep-breathing exercises in a gamified fashion, with challenges and a sense of progression that are absent in PACE. At the onset of the study, however, we were concerned that a traditional paced-breathing intervention would *trivially* be less engaging than a video game, making the result a foregone conclusion. For this purpose, we added touch interaction and a motor entrainment mechanism to PACE to make the comparison against more GBF meaningful. It appears, though, that adding these simple interactive elements to a breathing exercise made the task as engaging as a game.

If not engagement, what could explain GBF's higher adherence to treatment? A possible explanation may lie in the changes in engagement ratings over time. When comparing engagement scores in the first and last sessions, we found that all measures (ease of remaining in focus, ease of reaching the target BR, and enjoyment) dropped significantly for PACE, but not for GBF. An alternative explanation may hide in participants' feedback at post-test. When asked about the most liked aspects of the apps, six GBF participants noted that the game was enjoyable, compared to no participants in the PACE group. Thus, social desirability bias [56] may have been at play when PACE participants were asked directly to rate how engaging the apps are, but not when asked (indirectly) to comment on what they liked the most about their app.

Taking the analyses above as a whole, we have evidence both in favor and against H2, which led us to *not reject* H2, although we cannot fully support it either.

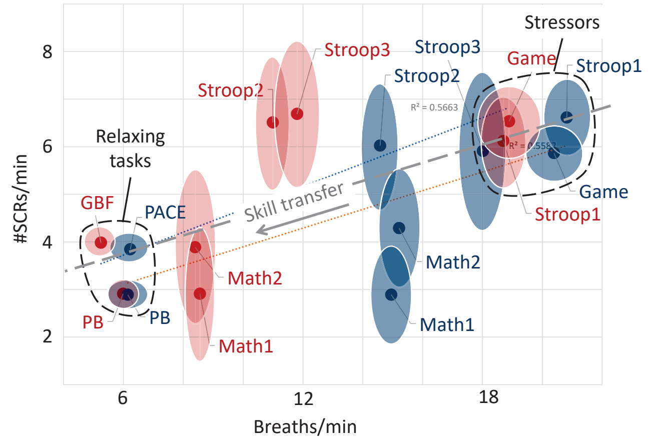


Fig. 8. Physiological response to all tasks in the study (ambulatory sessions are not included since EDA is not available for them). There is a strong positive correlation between BR and EDA, with skill transfer indicating a shift towards lower BRs and EDA.

5.2 Skill transfer (H3)

Our third hypothesis (H3) was that GBF would lead to better skill transfer, measured as lower BRs during a stressor. In fact, GBF participants obtained significantly lower BRs than PACE during post-test, supporting the three-day skill transfer hypothesis (H3b), the most challenging and critical of the two, as it shows learning effects that extend beyond individual sessions. However, GBF participants completed significantly more sessions than PACE participants, which raises the question of whether our skill transfer results were a dosage effect. If dosage was the main driver for increased skill transfer, and not the treatments' characteristics, we would predict a negative correlation between the number of sessions and BRs at post-test. However, our results (Supplementary Materials, Appendix 5) show no evidence of this: for GBF, correlations between dose and BRs during Stroop3/Math2 were not significant; for PACE, there was a significant correlation for Stroop3, but not for Math2. Thus, dosage effects do not explain why GBF led to better skill transfer.

Analysis of EDA shows mixed results. To recall, we found no significant reduction in EDA during Stroop2/3 compared to their levels during Stroop1. This would suggest that there was no skill transfer. However, this inference conflicts with the BR results, which show a significant drop from Stroop1 (pre-treatment) to Stroop2/3 (post-treatment), particularly for GBF. To reconcile the BR and EDA results, **Fig. 8** shows a joint plot of both physiological responses across tasks in the study, with the major and minor axes of the ellipses representing the standard deviation of the respective physiological variable. One can see a strong positive correlation between BR and EDA for both interventions ($\rho_{GBF} = 0.57$; $\rho_{PACE} = 0.56$), indicating that reductions in BR lead to reductions in EDA, for both interventions. The bottom-left shows a cluster of tasks designed to trigger a relaxation response at pre-test (low BR, low EDA). The top-right shows a second cluster of tasks designed to trigger a stress response at pre-test (high BR, high EDA). Connecting the two clusters is a diagonal line that shows the direction of skill transfer, along which the post-test stressors (Stroop2/3, Math1/2) are spread. Thus, it appears that GBF leads to skill transfer in terms of a *behavioral response* (voluntary changes in breathing rate), but not in terms of a *physiological response* (involuntary changes in EDA).

These mixed results are likely related to the unintended difficulty of Stroop2/3. To minimize learning effects, we reduced the response time from 3 sec for Stroop1 to 2 sec for Stroop2/3, which was hard to maintain for 5 min for a task whose average reaction time is 750ms [57]. Analyses of the Stroop scores corroborate this argument: both groups showed a significant drop in scores from Stroop1 (score ~ 90%) to Stroop2-3 (scores ~ 20-50%) – See Supplementary Materials, Appendix 6. Given such a dramatic reduction in scores, it is remarkable that GBF participants were able to reduce their breathing rates during CTW2/3 to half of what they had been during Stroop1.

5.3 In-the-moment relaxation (H4)

Hypothesis (H4) stated that PACE would outperform GBF in achieving in-the-moment relaxation. Participant ratings completed immediately before and after each ambulatory session show a significant improvement in positivity (valence) and relaxation (arousal) for the GBF group, but not for PACE, thus rejecting this hypothesis. At first, this result was puzzling: PACE is a simple task, with low cognitive load, whereas GBF is a more challenging task with high cognitive load (see NASA TLX in **Fig. 7a**). Thus, we expected that PACE would induce more in-the-moment relaxation than GBF. But physiological measures from the Laboratory Treatment sessions corroborate these self-report measures: GBF had significantly lower EDA than the Base Game and close to the EDA levels during the Paced Breathing training session. *In summary: adding respiratory biofeedback to the base game eliminated the latter's stress response or, alternatively, adding a gaming component to deep breathing did not eliminate the latter's relaxation response.*

5.4 Limitations and future work

A major challenge we faced in the present study was devising a meaningful way to compare two inherently different interventions. The GBF app requires eye contact and touch interaction, so comparing it against a control condition that would not require either seemed problematic. Also critical was to ensure that participants would perform deep breathing (i.e., the basic relaxation mechanism) for similar amounts of time in either intervention. Accordingly, we took a traditional paced-breathing exercise and modified it to include touch interaction and eye contact. This was not an arbitrary choice but a key decision that allowed us to (1) detect when participants stopped paying attention to the DB exercise, (2) alert them via a phone vibration, and (3) measure the total amount of *actual* practice time. The result of these design decisions was PACE. We took precautions to minimize disruptions. First, phone vibrations are brief (100 ms), so they are more of a subtle notification rather than an intense long vibration, which could be problematic. Second, phone vibrations are used sparingly: if the participant resumes tracking the bar within 5 seconds of the first vibration, a second vibration does not occur. Despite these precautions, when asked about negative aspects of the PACE app, some participants reported that it could cause frustration and stress. While the NASA TLX scores indicate that PACE is far less demanding than GBF, it is possible that some design elements in the PACE app may have biased the comparison against GBF. A potential solution to this issue is to allow participants to disable these features once they become proficient at deep breathing.

Both treatments aimed to encourage breathing at a target of 6 breaths/min. This choice was based on prior studies showing that HRV power is maximized at around 0.1Hz [58]. However, the optimal breathing rate (resonance frequency) varies from person to person. Thus, future work should explore whether customizing the target BR to each participant improves relaxation and skill transfer. To customize BRs, participants would undergo a calibration process, where they breathe at several rates, and the optimum is the one with the highest HRV power.

A potential direction of future work is to increase the variety of exercises, as several participants had suggested during the exit interviews. These could include different casual games, some designed to elicit excitement (e.g., Scale) and others designed to elicit contemplation (e.g., Flow [59], Monument Valley [60]), as well as apps that combine music with biofeedback [31]. A selection of such apps could be wrapped under a multi-armed bandit that would learn to provide recommendations for each participant. At first, participants would be encouraged to *explore* the various apps; over time, the system would provide recommendations to *exploit* the efficacy of various apps (e.g., reductions in BR, self-report measures). Offering a variety of apps could improve adherence and skill transfer, as participants would be able to practice deep-breathing skills in a variety of scenarios.

Our study required participants to wear a chest strap (BioHarness) that is unsuited for long-term use and broad adoption. Future work will explore additional methods to measure breathing rate that are less cumbersome. For example, most fitness trackers and smartwatches use photoplethysmography (PPG) to measure heart rate. Given the time series of heartbeats, it is possible to estimate breathing rates by detecting increases in R-R period during exhalation and decreases in R-R period during inhalation. Smartphone cameras could also be used to estimate eye gaze and facial expressions while participants practice their exercises. Eye-gaze would inform whether participants are focusing on the exercise or are distracted. Facial expression analyses could identify moments of boredom associated with low engagement. Smartphone cameras could also be used to estimate heart by examining subtle differences in skin color, as has been shown in non-contact image-based PPG [61].

Finally, a major next step of this research is to evaluate game-based biofeedback interventions in clinical populations, such as patients with anxiety disorders. These studies would examine whether game-based biofeedback can have beneficial effects in moments of crisis (as a coping behavior) and to lower overall anxiety (as a training procedure).

5.5 Conclusion

Our study indicates that gamified breath biofeedback (GBF) promotes better adherence to treatment, in-the-moment relaxation, and skill transfer than paced breathing. These findings indicate that GBF may be a good alternative self-regulation technique to complement traditional stress management interventions.

6 Acknowledgments

This work was supported by the National Science Foundation under Grant No. 1704636. We would also like to acknowledge the five anonymous reviewers, whose comments and suggestions helped improve the quality of this manuscript.

7 References

[1] S. Khalsa, "Yoga as a therapeutic intervention," in *Principles and Practice of Stress Management*, P. Lehrer, R. Woolfolk, and W. Sime Eds., 3 ed. New York, NY: The Guilford Press, 2007, ch. 17, pp. 449-462.

[2] R. Jerath, V. A. Barnes, D. Dillard-Wright, S. Jerath, and B. Hamilton, "Dynamic change of awareness during meditation techniques: neural and physiological correlates," *Front Hum Neurosci,* vol. 6, p. 131, 2012.

[3] D. R. Seals, N. O. Suwarno, and J. A. Dempsey, "Influence of lung volume on sympathetic nerve discharge in normal humans," *Circ Res,* vol. 67, no. 1, pp. 130-141, 1990.

[4] J. A. Cannon-Bowers and E. E. Salas, *Making decisions under stress: Implications for individual and team training*. APA, 1998.

[5] R. D. Rose *et al.*, "A randomized controlled trial of a self-guided, multimedia, stress management and resilience training program," *Behav Res Ther,* vol. 51, no. 2, pp. 106-112, 2013.

[6] K. M. Richardson and H. R. Rothstein, "Effects of occupational stress management intervention programs: a meta-analysis," *J Occup Health Psychol,* vol. 13, no. 1, p. 69, Jan. 2008.

[7] J. J. Gross, "Emotion regulation: Affective, cognitive, and social consequences," *J Psychophysiol,* vol. 39, no. 3, pp. 281-291, 2002.

[8] A. Kever, O. Pollatos, N. Vermeulen, and D. Grynberg, "Interoceptive sensitivity facilitates both antecedent-and response-focused emotion regulation strategies," *Pers Individ Differ,* vol. 87, pp. 20-23, 2015.

[9] M. G. McKEE, "Biofeedback: an overview in the context of heart-brain medicine," *Cleve Clin J Med,* vol. 75, p. S31, 2008.

[10] R. D. Croon, J. Geuens, K. Verbert, and V. Vanden Abeele, "A systematic review of the effect of gamification on adherence across disciplines," in *Proc. Int Conf Human-Computer Interaction*, 2021: Springer, pp. 168-184.

[11] H. Qin, P.-L. P. Rau, and G. Salvendy, "Effects of different scenarios of game difficulty on player immersion," *Interact Comput,* vol. 22, no. 3, pp. 230-239, Dec. 2010.

[12] D. Bormann and T. Greitemeyer, "Immersed in virtual worlds and minds: effects of in-game storytelling on immersion, need satisfaction, and affective theory of mind," *Soc Psychol Pers Sci,* vol. 6, no. 6, pp. 646-652, Apr. 2015.

[13] C.-H. Shih, N. Tomita, Y. X. Lukic, Á. H. Reguera, E. Fleisch, and T. Kowatsch, "Breeze: Smartphone-based acoustic real-time detection of breathing phases for a gamified biofeedback breathing training," *Proc ACM Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 3, no. 4, pp. 1-30, 2019.

[14] A. Parnandi, B. Ahmed, E. Shipp, and R. Gutierrez-Osuna, "Chill-Out: Relaxation training through respiratory biofeedback in a mobile casual game," in *Proc Int Conf Mobile Comput Appl Services*, 2013: Springer, pp. 252-260.

[15] A. Parnandi and R. Gutierrez-Osuna, "Visual biofeedback and game adaptation in relaxation skill transfer," *IEEE Trans Affect Comput,* vol. 10, no. 2, pp. 276-289, May 2017.

[16] A. Parnandi and R. Gutierrez-Osuna, "Partial reinforcement in game biofeedback for relaxation training," *IEEE Trans Affect Comput,* vol. 12, no. 1, pp. 141-153, June 2018.

[17] V. N. C. Gummidela, D. R. da Cunha Silva, and R. Gutierrez-Osuna, "Evaluating the role of breathing guidance on game-based interventions for relaxation training," *Front Digit Health,* vol. 3, Dec. 2021.

[18] T. Sonne and M. M. Jensen, "Chillfish: A respiration game for children with adhd," in *Proc 10th Int Conf Tangible, Embedded, Embodied Interaction*, 2016, pp. 271-278.

[19] A. Lobel, M. Gotsis, E. Reynolds, M. Annetta, R. C. Engels, and I. Granic, "Designing and utilizing biofeedback games for emotion regulation: The case of nevermind," in *Proc CHI (Extended Abstracts)*, 2016, pp. 1945-1951.

[20] S. T. Schlatter *et al.*, "Effects of relaxing breathing paired with cardiac biofeedback on performance and relaxation during critical simulated situations: a prospective randomized controlled trial," *BMC Med Educ,* vol. 22, no. 1, pp. 1-10, 2022.

[21] J. C. Brammer *et al.*, "Breathing biofeedback for police officers in a stressful virtual environment: challenges and opportunities," *Front Psychol,* vol. 12, p. 401, 2021.

[22] S. Bouchard, F. Bernier, É. Boivin, B. Morin, and G. Robillard, "Using biofeedback while immersed in a stressful videogame increases the effectiveness of stress management skills in soldiers," *PloS one,* vol. 7, no. 4, p. e36169, 2012.

[23] J. Weerdmeester, M. M. van Rooij, D. F. Maciejewski, R. C. Engels, and I. Granic, "A randomized controlled trial assessing the efficacy of a virtual reality biofeedback video game: Anxiety outcomes and appraisal processes," *Technol. Mind Behav.*, vol. 2, no. 2, 2021.

[24] K. T. Larkin, C. Zayfert, L. G. Veltum, and J. L. Abel, "Effects of feedback and contingent reinforcement in reducing heart rate response to stress," *J Psychophysiol,* vol. 6, pp. 119-119, 1992.

[25] J. R. Stroop, "Studies of interference in serial verbal reactions," *J Exp Psychol,* vol. 18, pp. 643-662, 1935, doi: 10.1037/h0054651.

[26] C. J. Golden and S. M. Freshwater, "Stroop color and word test," 1978.

[27] R. Yahav and M. Cohen, "Evaluation of a cognitive-behavioral intervention for adolescents," *Int J Stress Manag,* vol. 15, no. 2, p. 173, 2008.

[28] H. Al Osman, H. Dong, and A. El Saddik, "Ubiquitous biofeedback serious game for stress management," *IEEE Access,* vol. 4, pp. 1274-1286, 2016.

[29] A. Parnandi and R. Gutierrez-Osuna, "Physiological modalities for relaxation skill transfer in biofeedback games," *IEEE J Biomed Health Inform,* vol. 21, no. 2, pp. 361-371, Mar. 2015.

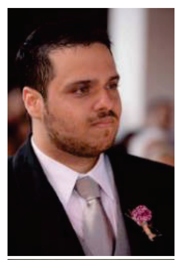
[30] B. A. Iwata, "Negative reinforcement in applied behavior analysis: An emerging technology," *J Appl Behav Anal,* vol. 20, no. 4, pp. 361-378, 1987.

[31] R. Bhandari, A. Parnandi, E. Shipp, B. Ahmed, and R. Gutierrez-Osuna, "Music-Based Respiratory Biofeedback in Visually-Demanding Tasks," in *Proceedings of the international conference on New Interfaces for Musical Expression*, Baton Rouge, Louisiana, USA, 2015, pp. 78–82.

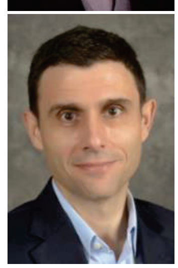
[32] G. M. Migliaccio, L. Russo, M. Maric, and J. Padulo, "Sports Performance and Breathing Rate: What Is the Connection? A Narrative Review on Breathing Strategies," (in eng), *Sports (Basel),* vol. 11, no. 5, May 10 2023, doi: 10.3390/sports11050103.

[33] M. H. Thaut, G. C. McIntosh, and V. Hoemberg, "Neurobiological foundations of neurologic music therapy: rhythmic entrainment and the

motor system," *Front Psychol,* vol. 5, p. 1185, 2015.
[34] R. Bechbache and J. Duffin, "The entrainment of breathing frequency by exercise rhythm," *J Physiology,* vol. 272, no. 3, pp. 553-561, Nov. 1977.
[35] D. M. Bramble and D. R. Carrier, "Running and breathing in mammals," *Science,* vol. 219, no. 4582, pp. 251-256, 1983.
[36] B. Raßler, D. Ebert, S. Waurick, and R. Junghans, "Coordination between breathing and finger tracking in man," *J Motor Behavior,* vol. 28, no. 1, pp. 48-57, 1996.
[37] P. Karthikeyan, M. Murugappan, and S. Yaacob, "A review on stress inducement stimuli for assessing human stress using physiological signals," in *Proc 7th Intl Colloquium Signal Processing Applications*, 2011, pp. 420-425.
[38] Y. Noto, T. Sato, M. Kudo, K. Kurata, and K. Hirota, "The relationship between salivary biomarkers and state-trait anxiety inventory score under mental arithmetic stress: a pilot study," *Anesth Analg,* vol. 101, no. 6, pp. 1873-1876, 2005.
[39] Oddrobo Software AB, "King of Math," vol. Android, 2019. [Online]. Available: https://play.google.com/store/apps/details?id=com.oddrobo.kom.
[40] Zephyr Technology. "Bioharness 3.0 User Manual." https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf (accessed 6/7/2023.
[41] Thought Technology Ltd. "Biograph Infinity Physiology Suite." https://thoughttechnology.com/content/docs/manual/SA7971%20V5.1%20Physiology%20Suite.pdf (accessed 6/7/2-2023.
[42] W. Boucsein, *Electrodermal activity*, 2 ed. New York, NY: Springer, 2012.
[43] R. S. Lazarus, J. C. Speisman, and A. M. Mordkoff, "The relationship between autonomic indicators of psychological stress: Heart rate and skin conductance," *Psychosom Med,* vol. 25, no. 1, pp. 19-30, 1963.
[44] T. M. Marteau and H. Bekker, "The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI)," *Br. J. Clin. Psychol.,* vol. 31, no. 3, pp. 301-306, 1992.
[45] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *J Res Pers,* vol. 41, no. 1, pp. 203-212, 2007.
[46] G. Strauss-Blasche, M. Moser, M. Voica, D. McLeod, N. Klammer, and W. Marktl, "Relative timing of inspiration and expiration affects respiratory sinus arrhythmia," *Clin Exp Pharmacol,* Aug. 2000.
[47] A. Parnandi and R. Gutierrez-Osuna, "Physiological modalities for relaxation skill transfer in biofeedback games," *IEEE J. Biomed. Health Inform.,* vol. 21, no. 2, pp. 361-371, 2015.
[48] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.
[49] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses," *J Behavior research methods,* vol. 41, no. 4, pp. 1149-1160, 2009.
[50] A. Parnandi and R. Gutierrez-Osuna, "Partial reinforcement in game biofeedback for relaxation training," *IEEE Trans. Affect. Comput.,* vol. 12, no. 1, pp. 141-153, 2018.
[51] T. Flenady, T. Dwyer, and J. Applegarth, "Accurate respiratory rates count: So should you!," *Australas Emerg Nurs J* vol. 20, no. 1, pp. 45-47, 2017.
[52] M. E. Dawson, A. M. Schell, and C. G. Courtney, "The skin conductance response, anticipation, and decision-making," *J Neurosci Psychol Econ,* vol. 4, no. 2, p. 111, 2011.
[53] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.,* vol. 39, no. 6, p. 1161, 1980.
[54] A. L. Gremaud *et al.*, "Gamifying accelerometer use increases physical activity levels of sedentary office workers," *J Am Heart Assoc,* vol. 7, no. 13, p. e007735, 2018.
[55] M. Dugas *et al.*, "Individual differences in regulatory mode moderate the effectiveness of a pilot mHealth trial for diabetes management among older veterans," *PloS one,* vol. 13, no. 3, p. e0192807, 2018.
[56] I. Krumpal, "Determinants of social desirability bias in sensitive surveys: a literature review," *Qual Quant,* vol. 47, no. 4, pp. 2025-2047, 2013.
[57] T. M. Sutton and J. Altarriba, "Emotion words in the mental lexicon: A new look at the emotional Stroop effect," *Ment. Lex.,* vol. 3, no. 1, pp. 29-46, 2008.
[58] E. G. Vaschillo, B. Vaschillo, and P. M. Lehrer, "Characteristics of resonance in heart rate variability stimulated by biofeedback," *Appl Psychophysiol Biofeedback,* vol. 31, no. 2, pp. 129-142, Julyy 2006.
[59] That game company. "Flower." https://thatgamecompany.com/flower/ (accessed 6/7/2023.
[60] ustwogames. "Monument Valley." https://www.monumentvalleygame.com/mvpc (accessed 6/7/2023.
[61] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.,* vol. 58, no. 1, pp. 7-11, 2010.



Dennis R. da Cunha Silva received a B.S. degree in computer engineering and a M.S. degree in systems engineering from the University of Pernambuco in 2012 and 2014, respectively, and a Ph.D. degree in computer science from Texas A&M University in 2023. His current research interests include affective computing, machine learning, and wearable sensors.



Ricardo Gutierrez-Osuna (Senior Member, IEEE) received a B.S. degree in electrical engineering from the Polytechnic University of Madrid in 1992 and M.S. and Ph.D. degrees in computer engineering from North Carolina State University, in 1995 and 1998, respectively. He is a Professor in the Department of Computer Science and Engineering at Texas A&M University. His research interests include speech processing, wearable sensors, and stress management.