Exploring the Impact of Gamification on BCI Performance in Children: The Case for Personalization

Dion Kelly  
University of Calgary

Brian Irvine  
University of Calgary

Eli Kinney-Lang  
University of Calgary

Daniel Comaduran Marquez  
University of Calgary

Erica D. Floreani  
University of Calgary

Adam Kirton (adam.kirton@albertahealthservices.ca)  
University of Calgary

Research Article

Keywords: brain-computer interfaces, pediatrics, electroencephalography, calibration, gamification, sensorimotor rhythms, motor imagery, P300 event-related potentials

Posted Date: May 30th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2987026/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

A major challenge with BCI use is the requirement for subject-specific calibration, which is often tedious and unengaging, but necessary to improve performance. This is especially true for children, whose limited attention and motivation may restrict the duration of endurable calibration periods. Several studies have shown that the addition of scoring systems and rewards to tasks, a process known as “gamification”, can increase motivation, attention, and task performance in children. This randomized, prospective, cross-over study aimed to address this challenge by comparing the effects of gamified versus non-gamified calibration environments on classification accuracy and BCI performance on utility-driven tasks.

Methods

Thirty-two typically developing children (14 female, mean age 11.9 years, range 5.8–17.9) attended two sessions lasting between 1.5-2 hours, to perform two standard paradigms: spelling using visual P300 event-related potentials (P300) and cursor control using sensorimotor rhythm (SMR) modulation, following gamified and non-gamified calibration. Gamified paradigms incorporated elements of game design, such as meaningful stories, quests, points and sounds. The primary outcome was BCI performance, which included performance of the classification model and online accuracy. Motivation, tolerability, and mental workload (NASA-TLX) were evaluated following each paradigm.

Results

For the P300 paradigm, mean classification accuracy was similar after gamified (96.81 ± 3.46%) and non-gamified (96.52 ± 2.42%) calibration. Mean classification accuracy for the SMR paradigm was 61.81 ± 13.35% with gamification and 59.84 ± 11.36% without gamification (n.s.). Mean online accuracy for SMR cursor control was 63.23% for both conditions. For the P300 spelling task, online performance was significantly lower following gamified training (p < 0.01). There were no significant differences found between classification accuracy, online BCI performance, motivation, tolerability, or perceived mental workload.

Conclusion

To our knowledge, this is the first study to investigate the effects of gamified calibration paradigms on classification accuracy and BCI performance in children. Our results reinforce the ability of typical children to control advanced BCI systems with performance comparable to adults. Gamified calibration environments may not enhance BCI classification and performance in children though the gamified
environments utilized in this study may not have been engaging enough. This work underscores the need for further research to optimize BCI training paradigms for pediatric use.

**Background**

Brain-computer interfaces (BCIs) have gained significant attention in recent years due to their potential to transform the lives of individuals with significant disabilities, enabling them to interact with their surroundings and communicate (1). BCIs function by detecting and translating brain signals into actionable commands, bypassing the need for physical interaction. Despite the promising advancements in BCI technology, a major challenge remains in the need for subject-specific calibration. This training is essential to ensure that the BCI system is tailored to the unique brain patterns and cognitive processes of each user, enabling efficient and accurate command execution. However, calibration can be a lengthy and tedious process, often requiring users to engage in repetitive tasks that may lead to decreased motivation and attention (2). This challenge is further exacerbated in pediatric populations, who typically have shorter attention spans and may find traditional BCI training protocols unengaging. As such, addressing the challenges of calibration is critical in optimizing BCI use and unlocking its full potential for various populations, including children who have been neglected from BCI research.

Implementing BCIs for children presents a unique set of challenges. Firstly, the pediatric brain is constantly undergoing development, with evolving cortical organization and changing neurophysiology (3). Much of this occurs in the face of major brain injuries early in life in target clinical populations such as youth with quadriplegic cerebral palsy. This complex neural plasticity complicates the process of designing and adapting BCI paradigms, as the algorithms and techniques optimized for adult brains may not be directly transferable to children (4). Additionally, children often have shorter attention spans and may require more engaging and motivating tasks to maintain their focus throughout the BCI training sessions. Thus, child friendly BCI paradigms must be developed that effectively capture and sustain user interest. Furthermore, children with disabilities who could immensely benefit from BCIs, may experience unique cognitive or sensory impairments that require specialized approaches to BCI design and training (5). Addressing these challenges is crucial to harness the potential of BCI technology to enhance the lives of children with disabilities and enabling them to achieve a greater degree of autonomy and independence (6).

Gamification represents one promising solution to improve BCI systems for pediatric users by making the training process more engaging and enjoyable. By incorporating elements of game design, such as points, quests, and meaningful stories, gamification can help maintain interest and motivation throughout training sessions (7). This increased engagement can potentially lead to better focus and sustained attention, which are crucial for effective BCI use. Moreover, gamification can create a sense of progression and achievement, further enhancing the intrinsic motivation of the child (7). By catering to the unique needs of children, gamified BCI systems can potentially reduce the barriers associated with traditional BCI training paradigms and improve overall performance. As a result, gamification could play an important role in bridging the gap between the potential benefits of BCI technology and its successful
implementation for pediatric populations, particularly for those with motor impairments or communication difficulties.

We have shown that both typically developing children (8, 9) and children with cerebral palsy (10–12) can learn to control EEG-based BCI systems using standard calibration paradigms. The aim of this study was to investigate the effect of gamified calibration on BCI performance on two common BCI control paradigms, the P300 event-related potential (ERP) and sensorimotor rhythms (SMR), in typically developing children. We hypothesized that gamified calibration would lead to improved performance in an online BCI classification task.

Methods

Participants

To ensure that our study was adequately powered to detect the effect of gamified calibration on BCI performance in children, we conducted an a priori power analysis using G*Power software (version 3.1.9.4). Based on pilot study results (13), we estimated an effect size of $f \geq 0.4$ for the interaction between calibration condition (gamified vs. non-gamified) and paradigm (P300 and SMR) in a 2x2 factorial within-subjects design. We set the desired statistical power ($1-\beta$) at 0.80 and the significance level ($\alpha$) at 0.05 for a two-tailed test. The power analysis indicated that a total sample size of $N = 31$ participants would be required.

Participants were recruited from the community via a database of typically developing children whose families have volunteered to participate in medical research (Healthy Infants and Children's Clinical Research Program)(14). Inclusion criteria were typical neurodevelopment with no neurological conditions, age 6 to 18 years, and informed consent/assent. Methods were approved by the Conjoint Health Research Ethics Board at the University of Calgary (REB21-1883). Participants received a $25 gift card for volunteering.

Protocol

We conducted a prospective, randomized, cross-over study to investigate the effects of gamified calibration in comparison to non-gamified calibration on two different paradigms: P300 and SMR. We employed a 2x2 factorial within-subjects design, with the factors being the paradigm (P300 vs. SMR) and the calibration condition (gamified vs. non-gamified). To ensure a balanced allocation of participants across the different conditions and to control for potential order effects, we utilized a Latin square design. Participants were randomly assigned to one of four conditions using an online randomization tool (www.randomlists.com). This resulted in four groups of eight participants each (total $N = 32$), with each group experiencing a unique sequence of the four combinations of paradigms and conditions. Participants attended two sessions lasting approximately 1.5 hours each. The primary outcome was classification accuracy, with secondary outcomes including precision and recall scores of the classification model from calibration, and the online accuracy achieved in subsequent BCI tasks.
Each visit had a similar protocol: 10-minute set up with a preliminary assessment questionnaire to record potential factors affecting performance (concussion, vision problems/corrected vision, mood, previous night's sleep, tiredness, exercise, food, alcohol/drug consumption including caffeine, and hobbies/sports), and the adapted Edinburgh handedness inventory (15). At the start of the first visit, participants also completed three tests from the CNS Vital Signs computerized neurocognitive test battery to evaluate attention and working memory: the Stroop Test, the Shifting Attention Test, and the 4-part Continuous Performance Test (16). At the start of the second session, participants completed two N-back tasks (1-back and 2-back) to evaluate working memory and visuospatial memory. The results from these neurocognitive tests will be reported in a separate paper.

Before beginning the BCI paradigms, the participants were instructed to sit still, with feet flat on the floor and hands/arms either resting on their lap or on the table in front of them. Participants were then evaluated on two classifier training environments for each paradigm: one gamified and one non-gamified. Afterward, they were assessed on their BCI performance on a “utility driven” BCI task: spelling for the P300 paradigm and cursor control for the SMR paradigm. After each paradigm, motivation, tolerability, and workload were evaluated with a series of questions.

The performance of the classification model was evaluated by calculating the metric “accuracy,” which represents the proportion of correct predictions among all predictions made. However, to provide a comprehensive assessment of the model's performance and ensure a balanced evaluation, we also reported precision and recall values alongside accuracy. Precision measured how many of the positive predictions made by the model were correct relative to the total number of positive predictions and serves as a measure of the model's ability to correctly identify positive instances without generating false positives. Recall measured how many of the true positive values in the test data were identified by the model relative to the total number of true values and helps us assess the model's sensitivity and its effectiveness in detecting all positive cases.

**Experimental Setup and BCI System**

A research-grade EEG-based BCI system was utilized for signal acquisition and processing (g.tec medical engineering GmbH, Schiedlberg Austria). Gamified and non-gamified classifier training paradigms were generated in-house.

EEG signals were acquired using the gSCARABEO active wet electrodes and amplified by the gUSB amplifier (g.tec medical engineering GmbH, Austria). Montages for SMR and P300 paradigms are specified in Table 1. All signals were sampled at 256Hz. Before initiating the BCI session, impedance was assessed to be < 5 kΩ for optimal signal quality and to minimize artifacts.

Table 1. Channel Configurations for P300 and SMR BCI Paradigms
Paradigm | # Channels | 10-20 Channel Locations
---|---|---
P300 | 8 | Fz, Cz, P3, Pz, P4, PO7, Oz, PO8
SMR | 16 | FC3, FCz, FC4, C5, C3, C1, CZ, C2, C4, C6, CP3, CP1, CPz, CP2, CP4, Pz

**BCI Paradigms**

One standard calibration scene, one gamified calibration scene, and one task scene were developed for each P300 and SMR (Fig. 1). For the gamified calibration scenes, we incorporated game elements such as a storyline and quest, scoring system, and audio-visual features like characters, background music, and sound effects associated with points.

**P300 Paradigm**

*Design*

A pseudo-random single character flashing design was chosen for the P300 paradigm due to its numerous reported advantages over the traditional row-column design, including mitigating adjacency-distraction effects and double-flashing errors (17). A 3x3 matrix (9 potential targets) was used for both calibration conditions and the spelling task, chosen to minimize total eye movement and user fatigue (17).

The familiar face paradigm is known to elicit stronger P300 responses in adults, leading to faster target selection and higher accuracy, in addition to increased potential for enhanced user engagement (18). These factors are crucial for optimizing BCIs for children with limited attention spans. Rezeika et al. (2018) suggested that using culturally well-known faces can lead to high and consistent effects across individuals (17); thus, we incorporated popular children's cartoon characters. Cells flashed one at a time, which means that only one key changes during each flash to display a cartoon character’s face in each flashing cell. Flashes lasted 100ms, with an inter-stimulus interval of 75ms.

*Calibration*

In the non-gamified calibration task, a target square was identified with a “+”. Users were instructed to pay attention to that square and silently count the number of times a character flashed, until the word “done” appeared on the screen, signaling the end of the training.

For the gamified P300-based BCI training, titled "Mole Patrol" (similar to whack-a-mole), participants were instructed to use their brain power to "whack" moles appearing from different holes. To achieve this, they had to silently count the number of times the mole appeared in the target hole, accompanied by a distinct sound, while ignoring other cartoon characters who were also attempting to catch the mole. Participants were guided through the game interface, with the number of moles "whacked" and the number of
remaining moles displayed in the top left corner, and the participant's score shown in the bottom right corner.

After completing a run, participants were required to enter the number of times the mole appeared in its hole. Correctly counting the mole's appearances earned them 100 points. The farther they were from the correct count, the fewer points they received for each mole. The highest possible score was 900 (9 runs x 100 possible points per run).

For both paradigms, each run consisted of 9 trials with a random number of ashes ranging from 10–15, which represented the number of times the mole emerged from its hole or the number of "ashes". The total calibration time was approximately 5 minutes. Once the classification model was trained, a confusion matrix was provided alongside percentages for the performance of the model, including the accuracy, precision, and recall.

Spelling Task

Using the classifier that was trained in the calibration phase, participants then completed a free spelling task using a two-stage T9 speller. Participants were required to write a five-letter word beginning at 10 flashes per key and decreasing by increments of two flashes for each subsequent word until two flashes per key, and a final word at 1 flash per key. Target words were different for each flash rate to mitigate potential learning effects. Thus, participants spelled a total of seven different five letter words. They were instructed to employ the same method as in the calibration task, by focusing on the key containing their target letter and silently counting the number of times the cartoon character flashed on the key. Each five-letter word required ten actions to spell - two actions (selecting a key and confirming the choice) for each of the five letters.

The accuracy of the task was calculated at each flash rate as the ratio of the correct number of selections (correctly chosen and confirmed keys) to the total number of selections (ten actions per each five-letter word). Participants were tested at 10 and 8 flashes. If their online accuracy dropped below 40% on two subsequent attempts, the task was discontinued to avoid causing frustration and disappointment.

SMR Paradigm

Design

For the SMR task, a binary class paradigm of left- and right-hand motor imagery was chosen. The training protocol was based off the Graz training paradigm, which is the standard training approach for motor imagery-based BCI (19). It involves the imagined movement of specific motor tasks, such as moving the left or right hand, while EEG signals are recorded and processed. After the first pair of imagined actions, real-time feedback was provided to users based on a classifier trained on the completed sets of imagined actions. The duration of the training protocol was 5.67 minutes with the following parameters: window length = 1.5s; number of training windows = 6; pause before training = 2s; number of training selections = 20; pause after training = 2s; train break = 4s.


**Calibration**

During the non-gamified calibration task, participants were instructed to concentrate on two gray squares presented on the screen. They were asked to imagine opening and closing the hand corresponding to the side with the larger square. The participants received feedback after one trial per side. The feedback was provided by making the square corresponding to the classified side more opaque. This real-time feedback encouraged participants to think more intently about the imagined hand movement and to continue imagining the movement until instructed to relax.

The gamified task, titled "Banana Dash," aimed to engage participants in teaching baby Moe, a virtual monkey character, how to catch falling bananas using their imagined hand movements. Participants were instructed to transfer their brain power to baby Moe by imagining grabbing the bananas with their left or right hand, depending on which side the bananas were falling. The task was designed to resemble a learning process, with Moe initially observing the participant's imagined actions before attempting to catch the bananas himself.

During the first two trials, Moe did not run towards the bananas, as he was simply observing the participant's thought patterns. After these initial trials, Moe would attempt to catch the bananas based on the participant's imagined hand movements. If Moe started moving in the wrong direction, participants were instructed to concentrate more intensely on imagining the correct hand movement to guide Moe towards the bananas.

Participants were encouraged to remain focused on consistently imagining the hand movements throughout the task and to keep imagining until Moe signaled them to relax. The objective of the game was to help Moe catch as many bananas as possible to achieve the maximum score. The gamified task aimed to maintain participant engagement and motivation while collecting motor imagery-based BCI data.

**Cursor Control Task**

The online task for the SMR paradigm was cursor control on a one-dimensional plane (horizontal) to answer yes/no questions by imagining left-hand and right-hand movements. The cursor control task entailed responding to ten yes/no questions with pre-confirmed answers. Participants were instructed to imagine left-hand or right-hand movements, which moved the cursor in the respective direction to indicate a 'yes' or 'no' response. Accuracy was calculated by dividing the number of correct responses by the total number of questions.

**Signal processing and Classification**

The classification steps for both P300 and SMR were identical in the standard and gamified calibration processes.

P300 Processing
After each visual stimulus the following 600ms of EEG data were recorded. Each 600ms epoch was filtered with a 5th order Butterworth bandpass filter with a passband of 0.1–15 Hz. Once the stimulus presentation for a single round was finished, the epochs for each object were ensemble-averaged, yielding one 600ms averaged epoch per object. These epochs were then saved with binary labels indicating whether the respective object was the target or not. This was repeated for each of the nine rounds of flashing. These epochs were then used to estimate the XDawn filtered ERP covariance matrices. ERP covariance matrices were mapped to their respective tangent space representations. This resulted in nine feature sets with the target label and 72 feature sets with the non-target label. To overcome this class imbalance, oversampling was used to oversample feature sets of the target class to match the non-target class. Feature sets were used to train a shrinkage linear discriminant analysis (sLDA) classifier using 5-fold cross validation. Reported classification accuracy did not include the oversampled feature sets used for training. This XDawn and sLDA were selected based on recommendations from Lotte et al. (20) for state-of-the-art BCI pipelines with small datasets. For the online spelling task, the trained classifier was used to evaluate each object’s posterior probability of belonging to the target class. The object with the greatest posterior probability was selected as the user’s chosen target.

SMR Processing

Each 9-second imagined action was segmented into 6 epochs of 1.5 seconds each. These epochs were filtered using a 5th order Butterworth bandpass filter with a passband range of 5–30 Hz. The covariance matrix was then calculated for each epoch. The covariance matrices were subsequently mapped to their corresponding tangent space representations, producing a feature set with a length equal to the number of channels. Logistic regression was employed for classification of the feature sets, following an approach similar to Barachant et al., but substituting logistic regression in place of LDA (21).

An iterative classification approach was chosen over a static one to provide real-time feedback. This feedback allowed users to adjust their mental strategies and enhanced the learning experience by offering more opportunities for improvement. A classifier was trained after the first two imagined actions, and updated after every subsequent pair of imagined actions, utilizing all available training data. Finally, a classifier was trained using all calibration data once it became available.

Questionnaires

Motivation was assessed using a shortened version of the Pediatric Motivation Scale (PMOT), which measures motivation as an event-based state from a child’s perspective (22). Workload was assessed using the child-adapted NASA Task Load Index (NASA-TLX), which measures subjective mental workload (23). The PMOT was developed for children as an event-based measure of motivation (rather than motivation as a trait) following activities. It contains 21 items which are divided into six subscales to evaluate subjective feelings of effort/importance, relatedness, autonomy, interest/enjoyment, competence, value/usefulness and open-ended items. Children responded to five items from the interest/enjoyment, effort/importance, and competence subscales using a 6-point ordinal face scale, where 1 was “not true at all” and 6 was “definitely true”. Tolerability was assessed at the end of each
session with a questionnaire using the same 6-point ordinal face scale. To promote valid responses throughout the scale, some items were framed negatively and were reverse scored. Children reported their perceived fatigue on a scale of 1 to 10 at the start of each session and at the end of each paradigm. To quantify the level of fatigue attributed to each paradigm, change in fatigue was calculated by subtracting the reported fatigue recorded after each paradigm from the reported fatigue recorded immediately before each paradigm.

**Data analysis**

Statistical analyses were performed using SPSS Statistics version 28.0 (IBM, USA), and GraphPad Prism version 9.0.0 (GraphPad Software, USA) was utilized for further statistical evaluations and the generation of graphs. Accuracy scores were averaged to obtain a mean score for each paradigm. Responses from questionnaires were transformed into corresponding numerical values. Outcomes were tested for normality using the Shapiro-Wilk test and parametric or nonparametric tests were used as appropriate.

To analyze the differences in accuracy across various flash rates for the P300 paradigm, we conducted a Friedman two-way analysis of variance (ANOVA) by ranks followed by pairwise comparisons using Dunn's Multiple Comparison's test. To report the overall online BCI accuracy for the P300 spelling paradigm, we computed the average accuracy score for runs within the stable range of flash rates where accuracy was not significantly impacted. Performance, encompassing classification model metrics (accuracy, precision, and recall) and online accuracy for utility-driven tasks, was compared between gamified and non-gamified calibration using Wilcoxon signed-rank tests. Differences in factors affecting performance, including motivation, tolerability, workload, and fatigue were analyzed using Wilcoxon sign-rank tests. Participants were stratified into four distinct age groups for the purpose of evaluating differences between ages. These groups were defined as follows: 'young' (6–8 years), 'middle' (9–11 years), 'young teens' (12–14 years), and 'teens' (15–17 years). A Kruskal-Wallis test was subsequently performed to evaluate potential differences in accuracy across these age groups. Bonferroni-Dunn corrections were applied to control for Type I errors in multiple comparisons.

A proficiency threshold for the SMR paradigm was set at 74%, which is the upper confidence limit of chance results for a 2-class classification model with 10 trials/class at $\alpha = 5\%$ (24, 25). This means that a classification accuracy higher than 74% would be considered statistically significant, indicating that the BCI performance is better than random.

**Results**

**Population**

The final population consisted of thirty-two typically developing children (14 female, mean 11.9 ± 3.9 years, range 5.8–17.9, 91% right-handed). There were no serious adverse events. Sessions averaged 60–90 minutes with the longest being 120 minutes. One participant dropped out after completing their first session. The participant who withdrew became discouraged and upset after comparing their performance
to that of their sibling and chose not to continue with the study. As a result, the final sample size was reduced to 31. For the P300 spelling task, one participant did not meet the threshold for task continuation (accuracy > 40%) in the first two runs (10 flashes and 8 flashes) and had a low recall score for their classification model. Given their low recall score, it was unlikely that their performance would improve in subsequent more difficult runs with lower flash rates. To prevent further frustration, the task was discontinued for this participant after the first two runs. Consequently, their data was not included in the calculation of the overall accuracy for the P300 spelling task.

Effect of gamification on BCI performance

Classification models

Mean performance metrics for the classification models (accuracy, precision, recall) are summarized in Table 1. The Wilcoxon signed-rank tests revealed no significant differences in classification model performance between the gamified and non-gamified calibration conditions for either paradigm for all three metrics: P300 accuracy (W = -50.0, p = 0.58); P300 precision (W = 17.0, p = 0.17); P300 recall (W = -67.0, p = 0.43); SMR accuracy (W = 110.0, p = 0.24); SMR precision (W = 121.0, p = 0.20); SMR recall (W = 38.0, p = 0.67).

In the gamified SMR calibration, 7 participants (22.58%) reached the 74% accuracy threshold, compared to 6 participants (18.75%) in the non-gamified calibration.
<table>
<thead>
<tr>
<th>Paradigm</th>
<th>Metric</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>P300 (non-gamified)</td>
<td>Accuracy</td>
<td>96.81%</td>
<td>3.46%</td>
<td>89–100%</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>94.61%</td>
<td>14.25%</td>
<td>50–100%</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>72.42%</td>
<td>28.17%</td>
<td>11–100%</td>
</tr>
<tr>
<td>P300 (gamified)</td>
<td>Accuracy</td>
<td>96.52%</td>
<td>2.42%</td>
<td>93–100%</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>99.00%</td>
<td>3.89%</td>
<td>83–100%</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>68.81%</td>
<td>20.83%</td>
<td>33–100%</td>
</tr>
<tr>
<td>SMR (non-gamified)</td>
<td>Accuracy</td>
<td>59.84%</td>
<td>11.36%</td>
<td>44–91%</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>59.61%</td>
<td>11.34%</td>
<td>45–92%</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>59.61%</td>
<td>12.95%</td>
<td>40–90%</td>
</tr>
<tr>
<td>SMR (gamified)</td>
<td>Accuracy</td>
<td>61.81%</td>
<td>13.35%</td>
<td>43–96%</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>61.94%</td>
<td>13.63%</td>
<td>42–98%</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>60.32%</td>
<td>15.05%</td>
<td>40–95%</td>
</tr>
</tbody>
</table>

**Effect of gamified calibration on subsequent performance on utility-driven tasks**

**P300 Spelling**

Figure 2 illustrates the mean online accuracy in the P300 spelling task across various flash rates, following both non-gamified and gamified calibration. After non-gamified calibration, online accuracy scores ranged from 82.19 ± 17.73% at 10 flashes (n = 32, range: 30–100%) to 38.89 ± 19.08% at 1 flash (n = 27, range: 0–90%). Conversely, after gamified calibration, scores varied from 74.84 ± 24.88% at 10 flashes (n = 31, range: 10–100%) to 42.27 ± 21.14% at 1 flash (n = 22, range: 10–90%). Significant differences in mean accuracy were observed at different flash rates, both after non-gamified ($\chi^2(6) = 94.62, n = 27, p < 0.0001$) and gamified calibration ($\chi^2(6) = 66.61, n = 22, p < 0.0001$). The flash rate's reduction, which amplified task difficulty, led to significant accuracy differences, especially noticeable when the flash rate fell below 4. Table 2 summarizes the results of the pairwise comparisons. Therefore, mean online classification accuracy reported for P300 spelling below is a combined mean accuracy for flash rates 4 through 10.
Table 2
Dunn’s Multiple Comparison’s test of accuracy between different ash rates for non-gamified and gamified conditions

<table>
<thead>
<tr>
<th>Flash rate</th>
<th>Non-gamified</th>
<th>Gamified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 vs 3</td>
<td>p = 0.034</td>
<td>n.s.</td>
</tr>
<tr>
<td>1 vs 4</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>1 vs 6</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>1 vs 8</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>1 vs 10</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>2 vs 4</td>
<td>p = 0.018</td>
<td>p = 0.020</td>
</tr>
<tr>
<td>2 vs 6</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>2 vs 8</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>2 vs 10</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>3 vs 6</td>
<td>p = 0.018</td>
<td>n.s.</td>
</tr>
<tr>
<td>3 vs 8</td>
<td>p = 0.009</td>
<td>p = 0.022</td>
</tr>
<tr>
<td>3 vs 10</td>
<td>p = 0.004</td>
<td>p = 0.007</td>
</tr>
</tbody>
</table>

Table 3 summarizes the mean online classification accuracy for both BCI tasks, P300 spelling and SMR cursor control, following gamified and non-gamified calibration. Mean online classification accuracy for the P300 spelling following non-gamified calibration group was 80.47 ± 17.07%, with a range from 45–100%, while accuracy following gamified calibration was 71.47 ± 19.93%, with a range from 18–100%. This median decrease in accuracy of 10% for the gamified condition was significantly lower compared to the non-gamified condition (W = -238.0, adjusted p = 0.017). Mean online classification accuracy for SMR cursor control follow non-gamified training was 63.23 ± 16.00%, with a range from 40–100%, while mean accuracy following gamified calibration was 63.23 ± 18.87%, with a range from 30–100%, which was not significantly different (W = 6.000, p = 0.90) (Fig. 3).
Table 3
Mean online accuracy for each utility driven task (P300 spelling and SMR cursor control) using the classification model generated in each condition (standard and gamified).

<table>
<thead>
<tr>
<th>Task (calibration condition)</th>
<th>Mean online classification accuracy</th>
<th>Std. Deviation</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>P300 Spelling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-gamified</td>
<td>80.47%</td>
<td>17.07%</td>
<td>45–100%</td>
</tr>
<tr>
<td>Gamified</td>
<td>71.47%</td>
<td>19.93%</td>
<td>18–100%</td>
</tr>
<tr>
<td>SMR Cursor Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-gamified</td>
<td>63.23%</td>
<td>16.00%</td>
<td>40–100%</td>
</tr>
<tr>
<td>Gamified</td>
<td>63.23%</td>
<td>18.87%</td>
<td>30–100%</td>
</tr>
</tbody>
</table>

Factors Affecting Performance

Age

No significant differences in accuracy were found for non-gamified calibration in the SMR paradigm, or for gamified calibration in either paradigm. Conversely, differences in accuracy were observed between age groups for non-gamified P300 calibration ($H(3) = 9.35, p = 0.025$), with teens achieving significantly higher accuracy compared to young children ($p = 0.026$).

Motivation

Motivation scores for the P300 paradigms were $5.03 \pm 0.79$ (range: 3.22-6.00) and $4.91 \pm 0.65$ (range: 3.44-6.00) for non-gamified and gamified conditions, respectively. For the SMR paradigms, scores were $5.02 \pm 0.64$ (range: 3.11-6.00) and $4.96 \pm 0.63$ (range: 3.33-6.00), for non-gamified and gamified conditions, respectively. No significant differences in motivation scores were found between the paradigms.

Tolerability

Tolerability scores for the P300 paradigms were $4.71 \pm 1.06$ (range: 2.00–6.00) and $4.73 \pm 0.75$ (range: 2.67-6.00) for non-gamified and gamified conditions, respectively. For the SMR paradigms, scores were $4.83 \pm 0.74$ (2.83-6.00) and $4.66 \pm 0.88$ (2.33-6.00) for non-gamified and gamified conditions, respectively. No significant differences in tolerability scores were found between the paradigms.

Mental Workload (NASA-TLX)

NASA-TLX scores for the P300 paradigms were $32.02 \pm 17.02$ (1.67–65.83) and $33.84 \pm 15.87$ (7.50-59.17) for non-gamified and gamified conditions, respectively. For the SMR paradigms, scores were
36.55 ± 13.56 (13.33–70.83) and 39.32 ± 15.37 (10.00–73.33) for non-gamified and gamified conditions, respectively. No significant differences in NASA-TLX scores were found between the paradigms.

**Fatigue**

The average change in fatigue for the P300 paradigm was 0.48 ± 1.85 (range -4.0 to 6.0) for non-gamified calibration and 0.16 ± 1.43 (range -3.0 to 6.0) for gamified calibration. For the SMR paradigm, the change in fatigue was 0.52 ± 1.85 (range -2.0 to 5.0) for non-gamified calibration and 0.20 ± 1.11 (range -2.0 to 3.0) for gamified calibration. There were no significant differences in change in fatigue found between calibration conditions for either paradigm.

**Discussion**

This study investigated the effect of gamified calibration on BCI performance in typically developing children on two well-known paradigms: P300 and SMR. The only significant effect of gamification was found on performance in the P300 spelling task, where gamified calibration had a possible negative effect on performance. Age-related differences in accuracy were found only in non-gamified P300 calibration, with teens performing better than younger children. Overall, our results suggest that gamification may not be beneficial for everyone and an ongoing need to define personalized approaches in BCI calibration.

The lack of serious adverse events and the duration of sessions indicates that the paradigms (both gamified and non-gamified) were generally well-tolerated by our young participants. However, the withdrawal of one participant due to discouragement elucidates many important considerations for introduction of BCIs to naive users, particularly children. A BCI, like many technologies, requires time to learn and adapt to (26). A new user may not perform well initially, leading to discouragement. This emphasizes one of the greatest ethical challenges in BCI deployment: the need to manage expectations, and ensure that users, particularly children, understand that it's normal to struggle in the beginning (27, 28). This also highlights the potential psychological effects of using a BCI, such as frustration or feelings of inadequacy, which can be detrimental to BCI performance and adoption (29). It's important to consider these effects when designing and introducing BCIs for children where research experience remains limited. This may also reflect on participant selection and preparation processes. Improved screening and education methods may be beneficial to ensure that participants are prepared for the challenges of using a BCI and understand that initial struggles are part of the learning process.

There is a notable lack of literature on BCI performance in children, making the findings of this study particularly informative regarding BCI applications in children. We demonstrated that the classification model performance in the P300 paradigm was consistently high in both non-gamified and gamified conditions, with mean accuracies exceeding 96%, indicating stable classification accuracy across paradigms. Additionally, gamification appeared to improve precision, reaching 99% (SD = 3.89%) in the gamified condition, while reducing variability compared to the non-gamified condition (SD = 14.25%). Recall rates were also relatively high, at 72.42% (SD = 28.2%) for non-gamified and 68.81% (SD = 20.8%)
for gamified conditions. To date, only one other study has investigated visual P300 BCI classification in children, comparing the performance of various classification models (30). That study found similar classification accuracies across all classifiers, ranging from 62–64%, with precision and recall rates between 61.5–63.5% and 62–66%, respectively. In contrast, our study exhibited substantially higher classification model performance. This suggests that our approach might offer advantages in experimental design, paradigms, or data processing techniques. Moreover, our results demonstrate the reliability of P300 BCIs as a tool for BCI control in children, irrespective of the presence of gamified elements. The high performance of the classification model in our study implies that P300 BCIs could be effectively employed in various applications for children, including assistive technology, rehabilitation, and gaming. These encouraging results may inspire the development of child-friendly P300 BCI systems that address the specific needs and preferences of younger users.

The performance of the classification model in the SMR paradigm was noticeably lower than in the P300 paradigm, with mean accuracies around 60% for both the standard and gamified conditions. This suggests that the SMR paradigm may be more challenging for BCI control, again regardless of whether calibration is gamified or not. This observation is consistent with previous research in adults, which has documented that P300-based BCIs generally outperform SMR-based BCIs (31). Our prior study evaluating P300- and SMR-BCI performance in children mirrored this finding (9). In the gamified SMR calibration, 7 participants (22.58%) reached the 74% accuracy threshold, compared to 6 (18.75%) in the non-gamified calibration. This suggests that gamification may have a slight (if any) positive impact on achieving the proficiency threshold. Compared to our previous study, six participants (20%) attained the threshold (mean accuracy was 55%), demonstrating that the SMR calibration methods employed in the current study neither improved nor diminished performance (9). It's important to note that children in this study only had two sessions using the SMR paradigm, with the total time for each session—including both calibration and the cursor control task—averaging around 10 minutes. Given the complex nature of motor imagery, proficiency is not typically achieved immediately; it's an acquired skill that usually requires repeated practice to gain mastery (32). It is also important to note that compared to the previous study, which utilized the g.tec mindBEAGLE system (25), the current study had fewer trials (10 vs. 30 per class). However, our trials were broken up into segments of shorter durations (20 segments of six 1.5s-epochs vs 60 segments of 8s) with longer duration of breaks between segments (8s vs 2s). Therefore, the SMR paradigm in this study had 180s of imagined movement with 160s of breaks, whereas the previous study had 480s of imagined movement with 90s breaks, which contributed to a longer calibration period (9.5 mins compared to 5.67 mins). The modified calibration approach, with longer breaks between segments and a reduced overall calibration period, did not negatively impact performance, and thus may be more suitable for children, potentially reducing fatigue and maintaining engagement during the calibration process. These results highlight the importance of tailoring BCI paradigms to the specific needs of children, to improve usability and effectiveness. Further research is required to identify more effective strategies to enhance BCI proficiency among children.

We did not find significant differences in performance (accuracy, precision, recall) between the gamified and non-gamified calibration conditions for either the P300 or SMR paradigms. It is worth noting,
however, that the high mean classification accuracy of 96% for both non-gamified and gamified P300 calibration paradigms suggests that there is limited room for improvement, making it challenging to discern any substantial benefits of the gamified approach. This ceiling effect may have masked the true potential of gamification in enhancing BCI performance, as the already high baseline performance in the P300 paradigm might overshadow any possible advantages. Our findings of no effect of gamification on SMR calibration align with the results from Castro-Cros et al. (2020), who found no effect of gamification on classification accuracy in an SMR-BCI rehabilitation application (33). There are several important factors that may have contributed to this outcome, including individual interests and the design of the gamified paradigms. The effectiveness of gamification can vary depending on individual preferences and interests (7). The gamified elements might not have been appealing or motivating enough, thus not impacting performance. There are several psychological mechanisms through which gamification affects motivation, including personal relevance and competence (7). In the present study, the gamified environments were designed by the researchers without input from the target population. This may have led to a mismatch between the participants' preferences and the implemented game elements, diminishing the motivational effects of gamification. Further, gamified tasks should offer challenges that are well-matched to the participants' abilities. If the tasks were too easy or too difficult, it could have led to boredom or frustration, respectively, and negatively impacted their sense of competence. By involving the target users in the design process and tailoring gamified elements to individual interests, BCI developers and researchers can ensure that challenges are well-matched and meaningful elements are incorporated into calibration paradigms, which may enhance users' motivation and engagement with the tasks (7).

While a high accuracy of the classification model during calibration is generally desirable, it does not guarantee high accuracy during online BCI control. Factors such as signal variability, user fatigue, and the complexity of the BCI task can all impact the accuracy of online BCI control. This was observed in the P300 spelling task, with online accuracies falling approximately 15% and 20% with classifiers generated from non-gamified and gamified calibration, respectively. What was more surprising was the lower online accuracy for gamified classifiers compared to non-gamified classifiers. This difference may be attributable to the differences in the design of the calibration tasks. Specifically, the presence of auditory stimuli in the gamified paradigm likely had a significant impact on the brain signals produced by the participants. In the Mole Patrol game, the characters made sounds when they popped up on the screen, with the mole making a distinct sound. These auditory stimuli likely elicited an auditory P300 in the participants' brain signals, which differs from visual P300 response in terms of distribution, latency, and amplitude (34). However, during the utility-driven spelling task, there were no sounds present, and therefore, the auditory P300 component would not be expected. This discrepancy between the presence of an auditory P300 component in the gamified calibration and its absence in the utility-driven task could have led to classifiers that were less effective in the latter context, resulting in a lower online accuracy for the gamified classifiers. In order to clarify the observed differences in online accuracy between gamified and non-gamified classifiers, we plan to conduct a detailed EEG analysis, which will be reported in a
separate paper. These findings highlight the importance of considering the context in which BCIs are used and calibrated, as these factors can significantly impact the system's performance.

No significant age-related differences were found in the gamified P300 calibration, while such differences were observed in the non-gamified P300 calibration, which suggests a possible influence of the gamification on age-related performance. Gamification could have made the task more engaging or enjoyable for younger children, improving their attention and thus their accuracy. This would reduce the performance gap between younger and older children that was observed in the non-gamified version. Alternatively, it's possible that the gamified version was less engaging or even distracting for older children, decreasing their accuracy and thus bringing their performance closer to that of the younger children. However, this disparity could also be attributed to age-related differences in the neural and cognitive processes underlying the P300 response, including attentional capacity (33,36). Future research could explore whether the age-related performance differences observed in the non-gamified P300 calibration could be mitigated through tailored gamification designs suitable for each age group.

Gamification did not significantly affect motivation, tolerability, or mental workload as measured by the NASA-TLX in either the P300 and SMR paradigm. Participants reported similar motivation and tolerability scores for both gamified and non-gamified conditions across paradigms, indicating comparable levels of engagement and comfort. The mental workload, as indicated by the NASA-TLX scores, was also consistent regardless of gamification, suggesting that the cognitive demand of the tasks remains stable regardless of the presence of game elements. This indicates that gamification, while potentially offering other benefits, does not appear to have strong effects on these specific aspects of user experience.

This study offers valuable insights into BCI use and development in children, revealing that gamification may not be universally beneficial as we anticipated. Our findings emphasize the importance of understanding individual variability in BCI performance and the need for personalized approaches in BCI design and training. Furthermore, we demonstrate that most children can effectively operate a P300-based BCI with high classification and online accuracies, while many can achieve above-chance classification accuracies for SMR BCIs. These insights emphasize the importance of ongoing research and development in the realm of BCIs for children. By concentrating on strategies to optimize BCI performance, such as gamification, researchers and developers can create more effective and accessible technologies to assist children with diverse needs, including those with motor impairments or communication challenges.

Several important limitations may have influenced our findings. The presence of sounds in the gamified P300 paradigm, for instance, could have introduced an auditory P300 component that was not present in the utility-driven task, potentially leading to a discrepancy in brain signal patterns between the two contexts. This might explain the lower online accuracy observed for gamified classifiers and highlights the importance of ensuring consistency in experimental conditions. The ceiling effect observed in the P300 paradigm may have concealed the true impact of gamification on BCI performance. Since the baseline performance was already high, the gamified paradigm might not have been able to demonstrate
significant improvements, which could undermine the potential benefits of gamification. This suggests the need for more sensitive measures or alternative paradigms that allow for a clearer assessment of gamification effects. In light of these limitations, future studies should consider maintaining consistency in experimental conditions, exploring alternative paradigms to overcome the ceiling effect, and increasing sample sizes for a more comprehensive understanding of individual differences. Engaging the target population in the design process to create personalized gamified elements may also enhance the effectiveness of gamification in BCI calibration, ultimately leading to improved performance and user experience.

Conclusion

This study contributes to the understanding of the impact of gamification on BCI performance in children and highlights the need for personalized approaches in BCI calibration. We demonstrated that while gamification may enhance BCI performance in some individuals, it may not be beneficial for everyone, indicating that a one-size-fits-all approach to BCI calibration is insufficient. By analyzing the effects of age, motivation, tolerability, and workload on BCI performance, this study has identified potential avenues for the development of targeted strategies to optimize performance in diverse user groups, including children. Future research is necessary to investigate individual variability and the need for tailored BCI design and training.

Abbreviations

BCI: Brain-computer interface; EEG: Electroencephalography; P300: P300 event-related potential; ERP: event-related potential; SMR: sensorimotor rhythms; sLDA: shrinkage linear discriminant analysis; PMOT: Pediatric Motivation Scale; NASA-TLX: NASA Task Load Index

Declarations

i. Ethics approval and consent to participate

Ethics approval for this study was granted by the University of Calgary Conjoint Health Research Ethics Board (CHREB). Ethics ID: REB21 – 1883.

ii. Consent for publication

Not applicable

iii. Availability of data and materials

The dataset supporting the conclusion of the article is available from the corresponding author on reasonable request.

iv. Competing interests
Authors DK, EKL, EDF, and AK are co-founders of Possibility Neurotechnologies, a start-up company developing personalized BCI solutions for children with disabilities. None received any compensation for the work submitted, and the company played no role in the study design, execution, or interpretation of results. The authors declare no other competing interest.

v. Funding

The Alberta Children Hospital Research Foundation and the Canadian Institutes of Health Research Funded this study.

vi. Authors’ contributions

DK conceptualized the research question and study design, recruited participants, collected, and analyzed all data and composed the manuscript. BI assisted with technology development and troubleshooting and data analysis. EKL assisted with concept formation, technology development, and troubleshooting. DCM assisted with technology development. EDF assisted with technology development. AK was the supervisory author and was involved with concept formation and study design and guided data analysis. All authors contributed to manuscript edits.

vii. Acknowledgements

We express our thanks to the children who participated in this study, and their families, for their time, patience, and invaluable contribution to this research.

References


Figures

Figure 1

Calibration and user task scenes. Top: P300 scenes — standard calibration scene (left), gamified calibration scene (Mole Patrol game, middle), T9 speller scene for spelling task (right). Bottom: SMR scenes — standard calibration scene (left), gamified calibration scene (Banana Dash game, middle), cursor control scene for yes/no response task (right).

Figure 2

Non-gamified
Gamified
Mean online accuracy for P300 spelling at different flash rates following non-gamified and gamified calibration.

Figure 3

Mean Online Accuracy for P300 Spelling and SMR Cursor Control Paradigms following Non-gamified and Gamified Calibration.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx