MoodSense Quiz — Project Report

Group 7 – Carlos Silva & João Paulo Pereira

1 Scientific question

Multiple-choice quizzes are common in education and casual learning apps, but they often treat all users the same: difficulty is fixed, feedback is uniform, and the interface rarely reacts to how the user is feeling. This can be a problem because short "failure streaks" can quickly turn into frustration and disengagement, especially when the user perceives the questions as unfairly difficult or when they believe they are performing worse than everyone else. MoodSense Quiz explores whether a quiz interface can support users more effectively by reacting to two signals: implicit adaptation based on user struggle and anonymized information about how previous users performed on each question.

The scientific question addressed by this project is:

How do streak-triggered hint system and anonymous "social ghost" cues (e.g., "X% of users answered correctly") influence user experience (perceived support, motivation, frustration) and performance in a short quiz, compared to a basic non-adaptive quiz interface?

We investigated whether a lightweight affective mechanism, based on quick self-report rather than emotion recognition, can increase the feeling of being guided, and whether social-ghost cues influence confidence and decision-making during difficult questions. We also examined whether these mechanisms translate into measurable performance differences (scores) in a short quiz session.

2 Related work overview

This project builds on three relevant research directions.

First, our work is grounded in gamification, commonly defined as the use of game design elements in non-game contexts. This framing is widely used to describe how "game-like" mechanics (feedback, progression cues, supportive interventions, etc.) can shape engagement and persistence. Deterding et al. propose the now-standard definition emphasizing "game design elements" applied outside games. At the same time, large reviews of empirical gamification studies report mixed outcomes: gamification can improve motivation and engagement, but effects vary with context, users and design choices rather than being universally positive. This motivates our decision to prototype specific mechanisms (mood adaptation + social cues) instead of relying on generic points/badges systems.

Second, the project draws from affective computing and emotion-aware human–computer interaction. Picard's work frames affect as important for natural interaction and decision-making and motivates systems that adapt to users' emotional states rather than treating emotion as irrelevant noise. In practice, emotion-aware systems can be built through physiological sensing, facial recognition, or self-report. For this prototype, we intentionally used self-report mood prompts because they are low-cost, transparent, and feasible in a classroom-scale study without complex sensing pipelines.

Third, our "social ghost" feature is inspired by social comparison and descriptive norms. Social comparison theory argues that people evaluate their abilities and opinions by comparing themselves with others when objective standards are not available. Similarly, normative conduct

research distinguishes between injunctive norms ("what people approve") and descriptive norms ("what people actually do") and shows that making a norm salient can shape behaviour. In our quiz context, the interface does not show identities or leaderboards; instead, it provides anonymous, aggregated cues such as how many prior users answered correctly, aiming to create a mild sense of context ("this was hard for others too") without triggering aggressive competition.

3 Proposed solution

We implemented two versions of the same quiz experience:

- **Basic Prototype:** The user answers a short set of multiple-choice questions. The interface provides standard correctness feedback after each question, but it does not change the quiz flow, difficulty, or options based on user state. No social information is displayed.
- **Emotion-Aware Prototype:** This version integrates two key mechanics:
  1. Streak-triggered hint system
  2. Social ghost cues (anonymous peer statistics)

Streak-triggered hint system

During the quiz, the system maintains a counter of incorrect answers When the streak reaches three, the system asks the user whether they would like a hint on the next question. The help is optional, allowing users to refuse assistance if they prefer to continue unaided. If the user accepts, the next question is modified by removing one randomly chosen incorrect option.

After the hint is applied, the wrong-answer streak is reset, preventing repeated hint prompting and re-establishing a neutral difficulty flow.

Social ghost cues

For each question, the prototype can show a small piece of aggregated performance information from prior attempts, more specifically the percentage of users who answered correctly. This is designed to affect the user's interpretation of difficulty ("this question is objectively less intuitive") and confidence calibration, while keeping the environment non-competitive and anonymous.

4 Results

We collected two types of results: objective performance logs and questionnaire feedback on user experience.

4.1 Performance logs

We analysed two performance log files representing 5 attempts per question, giving us the average score per question for our 50 questions. Because each question has 5 attempts, accuracy values move in 20% increments (0%, 20%, 40%, 60%, 80%, 100%).

To compare conditions, we computed overall accuracy as the total correct answers divided by total attempts across all questions. In the Basic condition, participants answered 162 out of 250 responses correctly, resulting in an overall accuracy of 64.8%. In the MoodSense condition, participants answered 171 out of 250 responses correctly, resulting in 68.4% accuracy. This represents a small improvement of 3.6% in favor of the MoodSense condition.

Figure 1 shows the percentage correct per question for both conditions. As you can see in the graphic, per-question accuracy varies substantially across the 50 questions, and the two conditions alternate in which one performs better depending on the item.

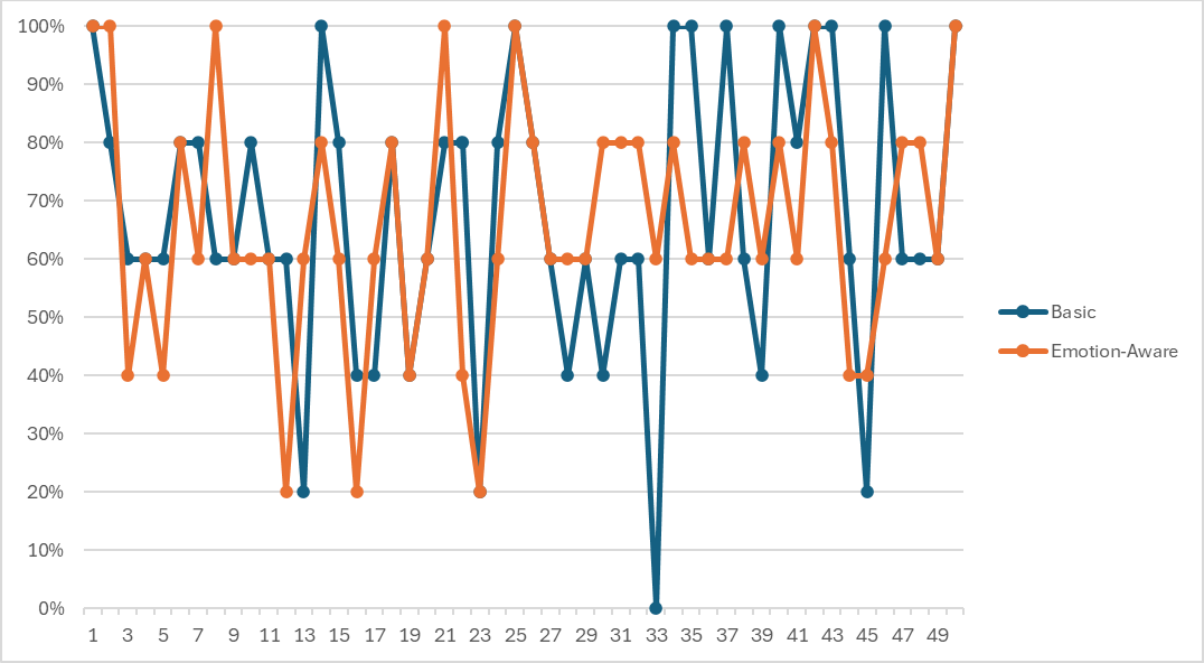

Figure 1: Percentage correct per question (Basic vs Emotion-Aware) across 50 questions.

Looking at individual questions, the MoodSense condition improved performance on some of the hardest items, but not consistently across all questions. For example, Question 33 shows a large shift: in the Basic condition it was answered correctly 0 out of 5 times, while in MoodSense it was answered correctly 3 out of 5 times. However, there are also cases where performance decreased, such as Question 30 (4/5 correct in Basic versus 2/5 in MoodSense).

4.2 Post-test questionnaire (user experience)

Ten participants completed a post-test questionnaire (5 tested Basic, 5 tested Emotion-Aware). For our scale we asked them to answer our questions on a scale from 1 to 7 in terms of agreement, and some other open-end questions.

The most striking difference between conditions was perceived support/guidance: users reported feeling much more guided by the interface in the Emotion-Aware prototype (mean $\approx$ 5.17) than in the Basic prototype (mean $\approx$ 1.25).

Perceived difficulty was lower in the Emotion-Aware condition (mean $\approx$ 3.83) than in Basic (mean $\approx$ 5.50).

Frustration ratings were similar across conditions (Basic $\approx$ 2.75 vs Emotion-Aware $\approx$ 2.50).

Enjoyment, however, did not clearly increase with adaptation (Basic $\approx$ 5.50 vs Emotion-Aware $\approx$ 4.67).

4.3 Qualitative feedback highlights

Open-ended responses provide insight into how the mechanisms were perceived:

- At least one Emotion-Aware user explicitly described using hints and noted that the interface reduced the number of options ("3 options instead of 4").
- Another Emotion-Aware participant reported that seeing the percentage of correct answers influenced their thinking, helping them judge whether a question was "not so obvious" and reducing overthinking.

- Several Basic users reported either guessing when stuck or feeling that some questions were "a little too niche," suggesting that without adaptive support, difficulty can feel unfair rather than simply challenging.

5 Discussion

MoodSense Quiz explored whether two lightweight interface mechanisms—anonymous "social ghost" statistics and an opt-in hint triggered by repeated errors—can improve the quiz experience and performance compared to a basic quiz. Instead of directly asking users how they feel, the system treats a streak of wrong answers as a proxy for struggle: after three mistakes, it offers a hint for the next question, which the user may accept or refuse. This keeps the user in control while still providing support when needed. In parallel, each question displays how often previous users answered it correctly, aiming to contextualize difficulty without creating competition.

Overall accuracy showed a small improvement in the MoodSense condition compared to the basic version. While this direction is consistent with the idea that contextual cues and optional assistance can help users recover from difficult moments, the magnitude is modest and should be interpreted carefully. The per-question pattern in Figure 1 is mixed: some questions improve (especially some difficult items), while others worsen, suggesting the mechanisms may help in specific struggle moments but do not guarantee uniform improvement across all content.

Despite modest performance gains, perceived support/guidance increased strongly in the Emotion-Aware condition, showing that adaptation and supportive UI cues were noticeable and interpreted as help rather than random changes. Lower perceived difficulty is consistent with the design goal: when users struggle, the system can reduce immediate difficulty (by removing one option), which makes the experience feel more manageable.

Frustration did not change substantially between conditions. In our case this may be related to pre-task mood reports: most users mentioned they were calm and focused, which meant that even our basic quiz did not induce strong frustration that the Emotion-Aware version could "fix."

Enjoyment did not increase with adaptation and was slightly lower in the Emotion-Aware version. One likely explanation is that interruptions (mood prompts) and visible intervention (help triggers) can reduce the "challenge satisfaction" that some users associate with quizzes, even while increasing support. Another possibility is that some participants experienced the quiz as "too easy" once adaptation kicked in, which reduces enjoyment for users who prefer challenge.

6 Limitations

- **Small sample size:** With only 10 participants total (5 per condition), we cannot draw strong or generalizable conclusions. Individual differences in skill, attention, and test-taking style could account for part (or even most) of the observed accuracy gap.
- **Between-subjects design:** Each participant completed only one version of the quiz, making the comparison more sensitive to individual variability. A within-subject or counterbalanced design would better isolate the effect of the MoodSense features.
- **Aggregated log format:** Our quantitative data came from logs that only store per-question totals (attempts and number correct). This prevented analysis at the individual level (e.g., how each user's performance evolved, who benefited most, and how hints changed subsequent behavior). In future work, we should ensure each user's data is logged separately (including time per question, hint acceptance/refusal, and streak events) to enable more accurate and diagnostic analyses.

7 Conclusion

In conclusion, our study suggests that combining difficulty-context cues with user-controlled assistance is a promising direction for making quizzes feel more guided while preserving autonomy. MoodSense Quiz demonstrated that simple, non-intrusive adaptations can be implemented without emotion recognition or competitive gamification. While performance gains were modest, the user experience results indicate that users felt substantially more supported and perceived the quiz as less difficult. Future work should increase the number of participants, use a within-subject or counterbalanced evaluation design, and collect per-user logs to better isolate the contribution of each feature and support stronger conclusions.

References

Deterding, S., et al. "Gamification: Toward a Definition Sebastian." Studies in Computational Intelligence 300 (2010): 1–361.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? - A literature review of empirical studies on gamification. In Proceedings of the 47th Annual Hawaii International Conference on System Sciences, HICSS 2014 (pp. 3025–3034). Article 6758978 IEEE.

Picard, Rosalind W. Affective computing. MIT press, 2000.

Festinger, Leon. "A theory of social comparison processes." Human relations 7.2 (1954): 117–140.

Cialdini, Robert B., Carl A. Kallgren, and Raymond R. Reno. "A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior." Advances in experimental social psychology. Vol. 24. Academ