Universiteit Leiden
The Netherlands

Data Science & Artificial Intelligence

The integration of facial expression recognition as a game mechanic in video games

Rachel Dijkstra

Supervisors:
Joost Broekens & Mike Preuss

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl 13/08/2023

Abstract

In this bachelor thesis, the use of facial expression recognition as a game mechanic to enhance the player experience in video games is examined. The study compares two versions of a dialogue-based game: one using traditional button-based dialogue selection and the other incorporating facial expression recognition. A pilot study compares three facial expression recognition models, highlighting the strengths and limitations of each. The main study employs a within-subject experimental design to assess user experience using self-report measures, including the System Usability Scale (SUS), Game Experience Questionnaire (GEQ), Positive and Negative Affect Scale (PANAS), and NASA Task Load Index (NASA-TLX). Contrary to the initial hypothesis, the facial expression recognition version receives lower usability scores, but no significant differences in overall experience. Participants express advantages such as ease of expressing emotions but note challenges related to changing expressions during reading and more usability issues. The conclusion suggests addressing these challenges and explores further research directions, including refining recognition models, integrating speech and audio features, and conducting longitudinal studies. This thesis provides insights into leveraging facial expression recognition as a game mechanic and contributes to understanding its impact on user experience in video games.

Contents

1 Introduction

Within contemporary society, video games have gained widespread popularity as a common form of leisure that numerous people enjoy for relaxation. As a widely practiced hobby, video games have attracted a considerable amount of attention in academic research lately. The video game industry constantly strives to increase customer satisfaction and sales by finding innovative ways to enhance the gaming experience.

Affective computing has emerged as a research area that offers potential for enhancing user experience within video games. This thesis aims to explore the possibility of using affective computing as a game-mechanic to enhance user experience in video games.

1.1 Motivation and related work

In this section, the existing literature and prior works related to this thesis will be discussed. This will include the extant research and implementations of affective computing in video games, as well as the various techniques employed to measure emotions in computer systems. Additionally, this section will include the techniques that have been employed to measure user experience within video games.

1.1.1 Affective computing in video games

Numerous studies have explored the integration of affective computing into video games, which involves using technology to detect and interpret human emotions. Affective gameplay, defined by K.M. Gilleade et al. [KGA05] as manipulation based on player affect, has been proposed to enhance engagement. Gilleade et al. [KGA05] proposed three design heuristics for affective gaming: "assist me", "challenge me", and "emote me" modes. The "assist me" mode uses emotion analysis to identify negative emotions, adapting the game to assist the player. The "challenge me" mode adjusts the game's difficulty level to become more challenging, enhancing engagement. Lastly, the "emote me" mode aims to provide emotional experiences and prevent emotional experiences from being reduced.

The aforementioned modes exemplify designs utilized in the development of adaptive affective games. B. Bontchev and D. Vassileva [BV17] have presented a game implementation in their study, where the difficulties in the "Rush of gold" game were dynamically adjusted. This was done by analyzing player performance and affection, which were measured through electrodermal activity and facial expressions. Likewise, in a study conducted by V. Vachiratamporn et al. [VVN14], an affective survival horror game was designed, and the affect of users was measured through their self-developed affect annotation tool (AAT). Two versions of the game were created, an affective and non-affective version, and were both evaluated. The findings revealed that there was no significant difference in the evaluation of the affective version and the non-affective version. They however believe that the game shows good potential for future research in the field of affective gaming.

In the third study conducted by Y. Izountar et al. [YIZ22], they introduced a virtual reality-based adaptive exer-game system called VR-PEER (Personalized Exercise and Emotion Recognition).

This system mainly focuses on emotion recognition. The authors utilized an emotion recognition module to examine the affective state of the player. In addition, they also utilized an adaptation module to modify the game based on the observed affective state. As a case study, the authors designed a serious game that was entirely virtual reality-based. For the emotion recognition module, facial expressions were utilized to determine the particular game that the player should play during their rehabilitation exercises.

In a separate study conducted by M. Granato and colleagues [MGR20], physiological data was collected in racing games on both a conventional monitor and a virtual reality (VR) headset. Machine learning techniques were applied to the acquired dataset to predict the emotions of players during game sessions. The findings of this study could be utilized to define game devices that have the ability to measure physiological data and improve game design.

These studies are merely a small subset of the abundant number of existing implementations and studies in the field of affective computing in video games. The majority of these implementations either modify the game in accordance with the user's affective state to improve the user's performance and/or level of engagement [KGA05] [BV17] [YIZ22] [VVN14] or utilize feedback to improve level design within the game [MGR20].

1.1.2 Automated emotion recognition in computer systems

A commonly employed technique in affective computing within video games is the integration of automated emotion recognition into the system. This can be accomplished through a variety of means, such as Electroencephalography (EEG) [GCP06] [GCP20], which entails the use of a headset with installed electrodes for measurement purposes. Additionally, the use of Electrocardiogram (ECG) signals [TDN19] [KRM10] - which are physiological signals utilized for the interpretation of electrical activity in the heart - can be used for automated emotion recognition. ECG signals can be measured through several means, including the use of a Wireless bio sensor RF-ECG [KRM10] or a Spiker-Shield Heart and Brain sensor [TDN19]. Galvanic Skin Response (GSR) / electrodermal activity (EDA) / skin conductance (SC) [VL07] [GCP09] [GCP06] is another viable method of automated emotion recognition, in which electrical parameters of the human skin are measured. Several more measurements based on physiological signals could be utilized for automated emotion recognition.

Speech recognition could also be utilized as a means to analyze emotions [ZB20] [IC12]. The advantage of employing this technique, as opposed to the previously discussed methods, is that it only necessitates a microphone and computer hardware, while some of the other methods require complex and often costly measurement systems. Facial expression recognition is another feasible method of recognizing emotions [DS16] that does not necessitate the use of expensive and complicated hardware. This technique merely requires a webcam and a computer. Similarly, body postures and gesture analyses could be employed for automated emotion recognition [SPC14], requiring only a webcam and a computer.

1.1.3 Measurement of user experience in video games

In the field of video games, a positive user experience can evoke positive emotions in players, leading to increased engagement and motivation to play. Numerous factors contribute to determining whether a player is having a satisfying experience. The ISO standards define user experience as the "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" [fS18a]. This includes "users' emotions, beliefs, preferences, perceptions, comfort, behaviors, and accomplishments that occur before, during and after use" [fS18a]. As emotions evoked in the user are also part of user experience, positive and negative affect could influence user experience as well.

The ISO standards also emphasize that "User experience is a consequence of brand image, presentation, functionality, system performance, interactive behaviour, and assistive capabilities of a system, product or service. It also results from the user's internal and physical state resulting from prior experiences, attitudes, skills, abilities and personality; and from the context of use." [fS18a]. Usability, defined as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"[fS18b], therefore seems to be a component of user experience.

According to the cognitive load theory, individuals possess limited cognitive resources that need to be efficiently allocated among concurrent mental tasks [Coo98]. Cognitive load relates to the amount of mental effort needed by the working memory when someone is performing an activity. When a system demands excessive cognitive load, usability is diminished [Coo98], indirectly impacting user experience. Because cognitive load influences usability, it indirectly influences user experience as well.

Additionally, research suggests that experiencing pleasure in games is closely linked to feelings of immersion and engagement [DH00]. The state of "flow," characterized by being both immersed and engaged, is associated with optimal experiences and profound enjoyment [Csi90]. Autonomy and competence have also been identified as factors contributing to positive experiences while using technology [Has08].

One commonly used self-report measure to assess positive and negative affect is the Positive and Negative Affect Scale (PANAS). [DWT88] This questionnaire utilizes a Likert scale format, where participants rate their experience of 10 positive emotions and 10 negative emotions on a scale of 1 to 5. A score of 1 indicates not feeling the emotion at all, while a score of 5 represents feeling the emotion to a great extent.

To evaluate usability, the System Usability Scale (SUS) [Bro95] is widely employed. This self-report measure comprises 10 questions that collectively assess the perceived usability of a system. Participants provide ratings on a Likert scale, ranging from Strongly Agree to Strongly Disagree.

The NASA Task Load Index (NASA-TLX) [HS88] is a self-report measure used to gauge cognitive load within a system. It assesses Mental Demand, Physical Demand, Temporal Demand, Own Performance, Effort, and Frustration Level. Participants rate each component on a Likert scale, indicating the level of each aspect from very low to very high. Scores for each component can range from 0 to 100 points.

The Game Experience Questionnaire (GEQ) [JHB09] is a self-report measure frequently used to assess player experience in games. It comprises 33 questions that evaluate competence, sensory and imaginative immersion, flow, tension/annoyance, challenge, negative affect, and positive affect. This questionnaire has gained widespread adoption in research, with over 64,549 downloads.

These measurement techniques offer valuable insights into different aspects of user experience in video games, encompassing emotional states, perceived usability, cognitive load, competence, sensory and imaginative immersion, flow, tension/annoyance, and challenge. By employing these measures, researchers can systematically evaluate and understand (part of) the user experience in video games.

1.1.4 Motivation

As previously discussed, research concerning affective computing integrated in video games typically comprises adapting game mechanics based on the predicted or observed states of players, or utilizing data or models that map the user's affective state to, for instance, improve video game design. However, the use of emotion recognition, such as facial expression recognition, as a game mechanic itself has not been studied often. Instead, it is primarily employed as a tool to adjust or analyze game mechanics. Therefore, the aim of this thesis is to examine the effect of utilizing facial expressions as a game mechanic on user experience.

2 Research question

The research conducted in this study aims to investigate the effect of automatically recognized emotions as a filtering mechanism in a game for the available dialogue options in a conversation with a non-player character (NPC) on player user experience in video games. This gives rise to the following research question:

What is the effect of automatically recognized emotions as a filtering mechanism in a game for the available dialogue options in a conversation with an NPC on user experience?

2.1 Hypotheses

Based on the literature review and the objectives of this research, the following hypothesis is proposed:

H1: The use of automatically recognized emotions as a filtering mechanism for dialogue options in conversations with NPCs will significantly improve user experience in terms of usability, cognitive load, emotional involvement (positive and negative affect), competence, sensory and imaginative immersion, flow, tension/annoyance and challenge compared to a traditional dialogue system without emotion recognition.

- H1a: Players in the experimental group (the group with emotion recognition) will demonstrate higher usability in the game world during the conversation with NPCs compared to the control group.
- H1b: Players in the experimental group will exhibit stronger positive affect in the game world during the conversation with NPCs compared to the control group.
- H1c: Players in the experimental group will demonstrate lower cognitive load in the game world during the conversation with NPCs compared to the control group.
- H1d: Players in the experimental group will improve on elements like competence, sensory and imaginative immersion, flow, tension/annoyance and challenge in the game world during the conversation with NPCs compared to the control group.

It is expected that the use of automatically recognized emotions in the dialogue system will provide players with more engaging and personalized conversational experiences. By filtering the available dialogue options based on recognized emotions, the system can dynamically adapt to the player's emotional state, possibly leading to improved user experience.

3 Method

In order to carry out this study, several tools needed to be developed and implemented. Firstly, an implementation of a facial expression recognition model was required to analyze facial expressions in frames. The decision to utilize facial expression recognition as a way to automatically detect emotions is primarily based on the idea that humans usually know better how to portray artificial emotions through facial expressions, than through other ways in which emotions could be automatically detected.

Since most models were built in Python, it was necessary to construct a Python API that incorporated some of these models. To determine the most suitable model for the game developed to address the research question, a comparative pilot study was designed to evaluate and compare the models. Additionally, a questionnaire was developed to assess the research question. The following sections discuss the development methodology for these tools.

3.1 API: facial expression recognition

The game and pilot study for this thesis were developed using JavaScript and Unity (C#). In order to incorporate facial expression recognition models, it was necessary to find a way to integrate them. One widely known framework for JavaScript is called faceAPI. However, the majority of facial expression recognition models were developed using Python. Therefore, an API was created with Python to allow access to various Python-based facial expression recognition models from any programming language. This approach significantly reduces the effort required to implement facial expression recognition in both the game and the pilot study.

The machine learning models utilized in this API were pre-existing models developed by other researchers. The MTCNN model from the facenet_pytorch package [Esl] was employed for face detection in frames. This model is based on the work of [KZQ16] and the implementation by [San][FSP15]. For facial expression recognition, two existing models were utilized. The 'model_filter.h5' model [Sha] and the 'enet_b0_8_best_afew.pt' model from the hsemotion.facial_emotions package [Sav] were used. The latter model is based on the research by [Sav21, Sav22, Sav23, AVSM22]. The models were selected by looking at the documentation and validating whether the statistics and analyses done by the creators showed promising results.

More information on the implementation of the API could be found in appendix E.

3.2 Pilot study: comparing facial expression recognition models

To determine the suitable model and facial expressions for the game designed to test the research question, a pilot study was conducted. The main aim of this pilot study was to compare the performance of three different models across multiple subjects. The objective was to assess and identify the model that demonstrated the most accurate grouping or clustering of emotions.

3.2.1 Materials

The pilot study involved the utilization of three facial expression recognition models. Two of these models were the same as those mentioned in the API section and were accessed through the API. The third model, known as faceAPI [Mü], was specifically designed for JavaScript. This model was the most commonly used one that could be found for Javascript.

The pilot study was conducted using a website developed with HTML and JavaScript. The website consisted of two main pages: an introduction page and a main page dedicated to the pilot study itself. The introduction page (Figure 1) provided instructions for participants and information regarding their privacy.

On the main page (Figure 2), there was a webcam view, a submit button, and text indicating the specific emotion participants were required to portray. To collect data, participants used a laptop or personal computer equipped with a webcam. To ensure ease of guidance, an external monitor was employed. The webcam captured frames at a resolution of 640 x 480, while both the monitor and laptop had a resolution of 1920 x 1080.

More information on the implementation of the pilot study can be found in appendix F

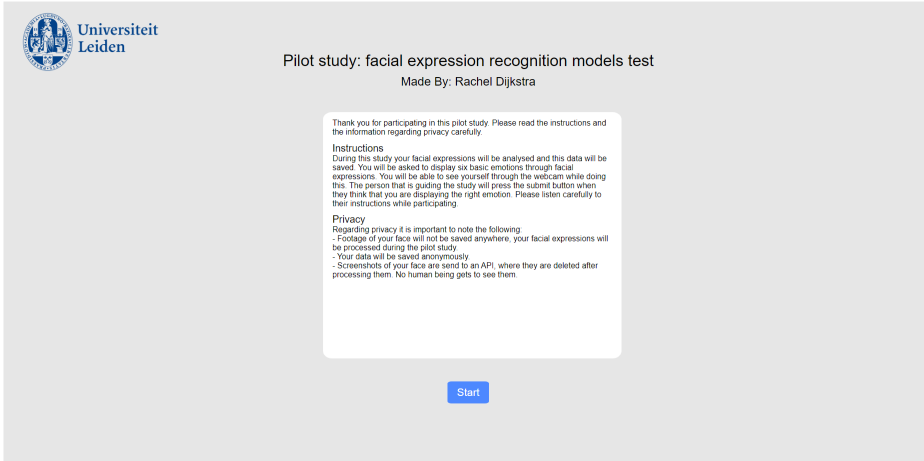

Figure 1: The introduction page of the pilot study, including instructions and information regarding privacy.

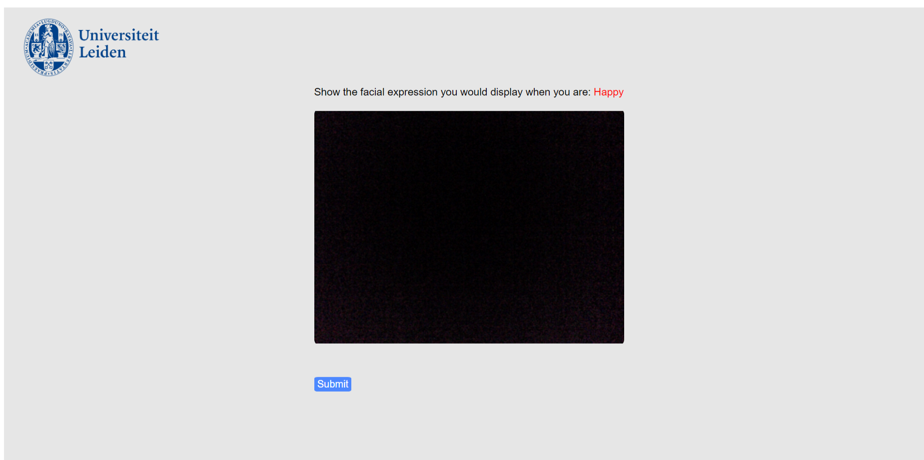

Figure 2: The main page of the pilot study. Here the pilot study itself is conducted. (black box is webcam view)

3.2.2 Experimental setup/approach

The experimental setup consisted of several key components and procedures designed to elicit and capture specific emotional expressions by different models. The following provides a comprehensive overview of the experimental setup:

Prior to engaging in the study, participants were directed to a webpage that presented instructions regarding the study's objectives and procedures. Additionally, the page contained information about the privacy measures implemented to safeguard the participants' personal information. It was mandatory for each participant to carefully read and agree to the terms before proceeding further.

After agreeing to participate, subjects were redirected to a separate webpage featuring a webcam interface. This page displayed a text prompt indicating the specific emotion the participant was required to portray. The participants were instructed to display six basic emotions: Happy, Sad, Angry, Fearful, Disgusted, and Surprised.

The researcher overseeing the study utilized a second monitor to guide the participants and initiate the data submission process. They closely monitored the participant's webcam feed and manually clicked the submit button on the website when the participant successfully conveyed the designated emotion. Following the activation of the button, participants were instructed to maintain a steady facial expression for approximately 5 seconds. Within this period, the facial expression recognition models analyzed 10 frames, and the resulting emotion probabilities for each frame were stored in JSON files.

The pilot study involved a total of 10 subjects. The participants' age ranged from 18 to 52 years. Among the participants, there were six male subjects and four female subjects. All participants were Caucasian. The only requirement for the subjects was that they were adults, as in theory, the models should work on any adult face.

Given the relatively small sample size of this pilot study, no statistical analysis was performed. It is important to note that the limited sample size may have implications for the generalisability and reliability of the study's results. It is crucial to acknowledge that the primary purpose of this pilot study was to select the model to be utilized in the game, rather than to provide substantial evidence regarding the superiority of any specific model.

3.2.3 Processing of acquired data

The data processing was conducted using Python. Separate Python files were created for each model. Within these files, the data from the JSON files of all participants for the specific model were loaded. For each emotion that could be predicted by the model, for each emotion that was tested during the pilot study, the mean probability was computed by aggregating the results from all participants. This process generated a table displaying the average predicted probability by the model for each emotion alongside the actual emotion portrayed by the participant. Subsequently, this table was transformed into a confusion matrix (heatmap).

The acquired data was processed using various Python packages. The Pandas package [comb] was employed to load the data into dataframes and facilitate data manipulation and analysis. The json package [Foub] was utilized to load the JSON files containing the data. To visualize the data as a confusion matrix/heatmap, the Matplotlib package [Dev] and the Seaborn package [Was] were employed. These packages provided the necessary tools and functions to create clear and informative visual representations of the data.

3.3 The game

In order to investigate the potential of using facial expression recognition as a game mechanic to enhance player experience in video games, two versions of a dialogue-based game were developed for this study. Recognizing the significance of facial expressions in communication, the decision was made to design the game as a dialogue interaction between the player and non-playable characters (NPCs).

The first version of the game followed a more traditional approach, employing buttons to filter and select dialogue options. In this version, the player would navigate through the dialogue choices using the provided buttons. The interface of a dialogue within this version can be seen on the left side of figure 3.3.

Contrastingly, the second version incorporated facial expression recognition as a novel mechanic. In this version, the player's facial expressions would determine the available dialogue options. The tone of the dialogue options presented to the player would be influenced by the facial expression exhibited on their own face. The interface of a dialogue within this version can be seen on the right side of figure 3.3.

Figure 3: The two interfaces of the dialogue within the game. On the left side is the dialogue interface of the traditional version. On the right side is the dialogue interface of the version with facial expression recognition.

By implementing both versions of the game, the aim was to compare the experiences of players utilizing the traditional button-based dialogue selection versus the facial expression recognition-based dialogue selection. Note that besides the difference in filtering the dialogue options, the two versions of the game are identical.

3.3.1 Game concept

Both versions of the game developed for this study fall under the RPG genre and feature an isometric view. The storyline revolves around the player assuming the role of a newly appointed tribe leader. Following the recent passing of their father and being an only child, the player finds themselves next in line for the leadership position within the tribe. However, the tribe holds little respect for the player, perceiving them as a privileged child. Consequently, the primary objective of the game is for the player to earn the respect of the tribe.

Achieving this goal entails interacting with tribe members and selecting appropriate tones and responses during these interactions. Additionally, there are various tasks or challenges, presented as mini-games, which can also contribute to earning the respect of the tribe members. Completing these tasks successfully can enhance the amount of respect for the player from the tribe. Conversely, choosing to give up on a task will result in an even bigger decrease in respect compared to not attempting the task at all.

The main mechanics implemented in the game involve the player navigating the isometric island using arrow keys or the 'WASD' keys to locate NPCs and tasks. Interactions with tribe members or tasks are initiated by pressing the 'I' key. In the traditional version, dialogue options and tone selection are facilitated through button clicks. In the version incorporating facial expressions, tone selection is achieved by choosing specific facial expressions. Additionally, a dedicated button allows the player to give up on a particular task.

The target audience for this game primarily consists of adults who have prior experience with dialogue-based games or RPGs, so they would be familiar with most of the game mechanics. This way, the facial expression recognition mechanic is the only novel element for players. It is worth noting that the target audience is adults due to the fact that facial expression recognition models are typically trained on adult faces, potentially leading to less accurate results when applied to children.

3.3.2 Game design

Various decisions were made to shape the visual and interactive elements of the game. The choice of an isometric view over a 3D view was driven by the need to prioritize performance, as a 3D view would require more graphical resources and could lead to stuttering in a webGL build. The webGL build was chosen to allow easy access to the game through a website.

The decision to use an isometric view was also motivated by the desire to create a more realistic look, resembling a 3D world. Given that the game's story revolves around a tropical tribe, the environment was designed as a tropical island. To match this theme, the game's interface incorporated tropical elements such as flowers, wooden buttons, and wooden dialogue boxes.

To save time and effort, AI tools like Midjourney and Leonardo.ai were used to generate certain elements of the game, including the main island, the mini-game island, and flower elements in the interface. However, the characters were hand-drawn to ensure consistency and to depict them as members of a tropical tribe. The decision to give characters large heads in proportion to their bodies was made to compensate for any odd proportions resulting from the AI-generated island and to maintain a cohesive art style.

Figure 4: The base interface within the game. This interface is for both versions the same.

The interface design included specific elements to enhance user experience. This included a respect bar in a red color, which is chosen to stand out, and serves to indicate the tribe's level of respect for the player. The text indicating the displayed emotion provided feedback to the user, allowing them to judge whether the game detected the right emotion and to adjust their expression accordingly. This text will also say "Unreadable" when the player's face is not positioned right. The menu button featured a recognizable icon commonly associated with menus, while the "give up" button used an X icon, commonly used to cancel actions, making it intuitive for players. Figure 4 shows the base interface of the game.

In the game design process, chatGPT was utilized to create the dialogue. Initially, the OCEAN characteristics, age class, gender, and job roles of the characters were determined for the dialogue. These factors were used to develop personas for the characters in collaboration with chatGPT. The detailed prompt used for this can be found in Appendix B.

Once the personas were created, situations for the dialogue were invented. In collaboration with chatGPT, specific dialogue options were then created. The prompt used for this can be found in Appendix C. The emotional tones or clusters, namely "Disapproval," "Distressed/Uncertain," and "Positive/Approval," were identified based on the findings of the pilot study. These three clusters were found to be easily recognizable through FACS-based emotion recognition in the study. These emotions guided the development of dialogue options.

To ensure a dynamic conversation, responses from the interacting character were also created in collaboration with chatGPT. The prompt for this can be found in Appendix D. The creation of the entire dialogue followed an iterative process involving these three steps. Each dialogue topic was carefully designed to connect with other dialogues, creating a cohesive narrative experience.

In terms of sound design, the focus for this game was relatively minimal. The game incorporated a background sound to evoke a tropical atmosphere. Additionally, roaring sounds were included to aid players in finding animals to hunt (task). Feedback sounds were implemented in the memory pairs mini-game to indicate when a card was turned and when a successful pair was found.

3.3.3 Development

The game development process involved working within Unity and utilizing C# scripting. Specifically, Unity version 2019.3.0f6 was used for this project. The decision to use an older version of Unity was driven by the lack of updated documentation on how to communicate with the Unity API, which grants access to specific functions within the Unity scripts of the game build. This functionality was necessary for integrating facial expression recognition.

To create the WebGL build of the game, Unity compiles the project into a folder containing an HTML file and several scripts.

In the version incorporating facial expression recognition, communication with the Unity API was required to access and trigger certain functions within the Unity C# scripts. To achieve this, a JavaScript script was written. This script enables access to the user's webcam (with their consent). Webcam frames are then sent to the facial expression recognition API every 150 milliseconds. The detected emotions are clustered into the emotional clusters recognized in the game. The emotional cluster with the highest probability is subsequently sent to the appropriate function in the Unity build through the unity API. The emotional state is displayed in the text located in the bottom left corner of the game, and during dialogues, the dialogue options are dynamically adjusted based on the recognized emotion.

The WebGL build can be accessed by clicking on the HTML file provided by unity after building it.

3.4 Main study: Comparison user experience

To assess the impact of automatically recognized emotions as a filtering mechanism for the available dialogue options in a conversation with a non-player character (NPC) on user experience, a measure needed to be developed to capture the users' experience while playing both versions of the game. Additionally, the collected data from these measures needed to be processed to facilitate interpretation. The following subsections describe the process undertaken to accomplish these objectives.

3.4.1 The development of the questionnaire

To assess user experience in both versions of the game, a custom measurement tool was developed. Given the subjective nature of user experience, self-report measures were deemed suitable for this purpose. The literature review identified several established measures for evaluating different aspects of user experience, including the System Usability Scale (SUS), the Positive and Negative Affect Scale (PANAS), the NASA Task Load Index (NASA-TLX), and The Game Experience

Questionnaire (GEQ). The specific questions from these measures can be found in Appendix G, H, I, and J, respectively.

To comprehensively capture various aspects of user experience, a composite questionnaire was created by combining questions from the aforementioned measures. The questionnaire included the same set of questions for both versions of the game, enabling a comparison of the results to identify potential differences in user experience. Additionally, respondents were asked an open-ended question: "What are your thoughts on using facial expressions as a game mechanic in the game that you've played?" This question aimed to capture users' experiences in their own words.

The questionnaire was developed using Google Forms and is available in its entirety in Appendix K.

3.4.2 Experimental setup/approach

To evaluate user experience in both versions of the game, a within-subject experimental design was employed. Participants were instructed to play both versions of the game and provide feedback using the developed questionnaire.

To minimize bias related to the novelty of the game, participants were divided into two groups. Three out of the five participants started by playing the traditional version of the game, while the remaining two participants began with the version incorporating emotion recognition. After completing the first version, participants switched to playing the other version. The intention was to ensure that each participant experienced both versions of the game.

To ensure a comprehensive assessment of user experience, participants were instructed to play each version of the game to completion. This approach aimed to provide sufficient exposure to the game, enabling participants to reliably answer the questionnaire.

Following the completion of both versions of the game, participants were provided with the questionnaire. They were requested to complete the questionnaire immediately after playing both versions, while their experiences were still fresh in their minds.

The game was played on a laptop equipped with a webcam capturing frames at a resolution of 640 x 480. The laptop screen itself had a resolution of 1920 x 1080.

The participant group consisted of five individuals, including four males and one female, with ages ranging from 17 to 51.

3.4.3 Processing of the data

To process the results obtained from the questionnaire, the data was initially exported to an Excel file. Within this file, the questions were categorized into sections based on the elements they measure. The scores provided by the participants were then subjected to normalization and transformation. This step ensured that higher scores consistently indicated a positive effect, while lower scores consistently indicated a negative effect. However, in the case of the NASA-TLX, higher scores were associated with a negative effect and vice versa.

To facilitate the analysis and comparison of results, the outcomes for each element were visualized using bar plots. This was done by using Python and the matplotlib package [Dev]. Each plot represented a specific questionnaire from the literature review, with each bar corresponding to a particular element within that questionnaire. By employing this visual representation, it became easier to identify any notable differences in user experience between the two versions of the game.

3.4.4 The development of a follow-up questionnaire

After analyzing the results obtained from the questionnaire, it was observed that certain responses yielded interesting findings. To further investigate the reasons behind these results, a follow-up questionnaire was developed.

The follow-up questionnaire was designed exclusively with open-ended questions. These questions focused on elements that contribute to user experience and asked questions about the reason behind variations in the experiences of both game versions. The aim was to gather qualitative insights and gather participants' perspectives on specific aspects of the games.

The complete follow-up questionnaire can be found in Appendix L.

4 Results

In this section, the results obtained from both the pilot study and the main study are presented. These findings are subsequently followed by a discussion section, where we delve into the implications and interpretations of the results.

4.1 Pilot study: comparing facial expression recognition models

This section presents the results obtained from the pilot study, focusing on the comparison of the three facial expression recognition models. The results are presented in the form of confusion matrices/heatmaps, which provide a visual representation of the model's performance in recognizing different facial expressions. The matrices are analyzed and compared to each other to assess the relative effectiveness and accuracy of the models in capturing and classifying facial expressions.

The confusion matrix/heatmap presented in the subsequent subsections can be interpreted as follows: The y-axis displays the actual emotions that participants were instructed to portray during the pilot study. On the x-axis, the predicted emotions by the model are shown. Each box within the matrix represents the predicted probability of an emotion on the x-axis corresponding to the actual portrayed emotion on the y-axis. An optimal performance would be indicated by a dark green diagonal line of boxes running from the top left to the bottom right of the confusion matrix. Such a diagonal line indicates accurate predictions, where the model successfully aligns the predicted emotions with the intended portrayed emotions.

4.1.1 faceAPI

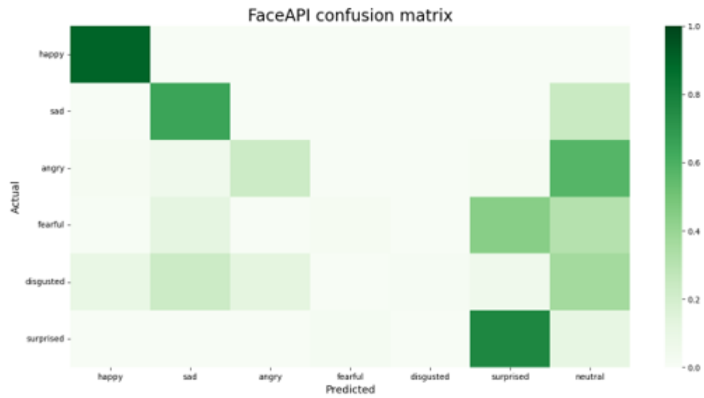

Figure 5: The confusion matrix of the results of the faceAPI model in the pilot study. The y-axis represents the emotions the participants had to display during the study. The x-axis represents the emotions predicted by the model.

Figure 5 depicts the results obtained from the faceAPI model, displayed in the form of a confusion matrix. In this matrix, we observe a dark green box within the 'Happy' row, indicating a strong correspondence between the predicted and actual 'Happy' emotions. This suggests that the faceAPI model demonstrates good recognition capabilities for the 'Happy' emotion, with a high predicted probability.

Furthermore, we observe relatively dark boxes within the 'Sad' and 'Surprised' rows, aligning with the corresponding predicted emotions. This implies that the model shows moderate success in recognizing these emotions, as indicated by the relatively high predicted probabilities.

Conversely, the 'Angry', 'Fearful', and 'Disgusted' emotions mostly result in neutral predictions, characterized by lighter boxes within the respective rows. Additionally, the 'Fearful' emotion exhibits a tendency to be predicted as 'Surprised'. These findings suggest that the faceAPI model struggles to accurately identify and differentiate these emotions, indicating limited recognition capabilities for 'Angry', 'Fearful', and 'Disgusted' emotions.

4.1.2 model_filter

Figure 6: The confusion matrix of the results of the 'model_filter.h5' model in the pilot study. The y-axis represents the emotions the participants had to display during the study. The x-axis represents the emotions predicted by the model.

Figure 6 displays the results obtained from the 'model_filter.h5' model, presented as a confusion matrix. Within this matrix, we observe a dark green box in the 'Happy' row, indicating a strong correspondence between the predicted and actual 'Happy' emotions. This suggests that the model recognizes the 'Happy' emotion well, which is indicated by the high predicted probability.

Additionally, we notice a relatively dark box in the 'Surprise' row, aligning with the corresponding predicted emotion. However, 'Fear' and 'Happy' also tend to be predicted instead of 'Surprise'. This implies that the model demonstrates moderate success in recognizing the 'Surprise' emotion, as indicated by the relatively high predicted probabilities.

In contrast, the 'Sad', 'Angry', 'Fear', and 'Disgust' emotions yield a range of predictions, reflected by lighter boxes within their respective rows. Specifically, the 'Sad' emotion displays a tendency to be predicted as either 'Sad' or 'Fearful', with similar probabilities. Similarly, the 'Angry' emotion is mostly predicted as 'Fear', with some predictions of 'Angry', 'Sad', and 'Disgust'. The 'Fear' emotion is predominantly predicted as 'Surprise' and 'Fear', suggesting a potential clustering of emotions between 'Surprise' and 'Fear'. Moreover, the 'Disgust' emotion is primarily predicted as 'Fear', with minor predictions of 'Happy', 'Sad', and 'Angry'. These findings indicate that the 'model_filter.h5' model encounters challenges in accurately identifying and distinguishing the 'Sad', 'Angry', 'Fear', and 'Disgust' emotions, showing limited recognition capabilities for these emotions.

4.1.3 enet_b0_8_best_afew

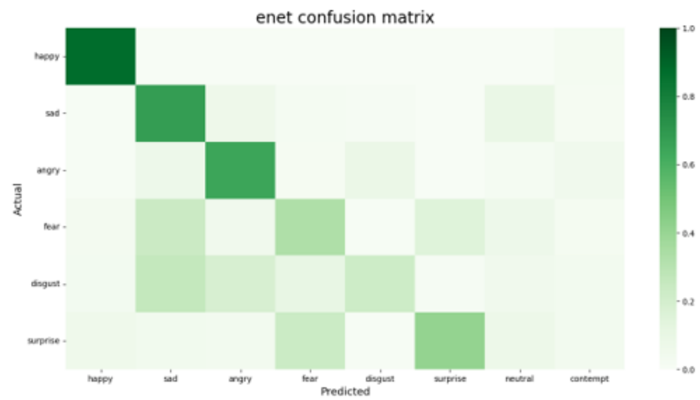

Figure 7: The confusion matrix of the results of the 'enet_b0_8_best_afew.pt' model in the pilot study. The y-axis represents the emotions the participants had to display during the study. The x-axis represents the emotions predicted by the model.

Figure 7 presents the results derived from the 'enet_b0_8_best_afew.pt' model, illustrated in the form of a confusion matrix. In this matrix, we observe dark green boxes within the 'Happy', 'Sad', and 'Angry' rows, indicating a noticeable correspondence between the predicted and actual emotions. This indicates that the model demonstrates strong recognition capabilities for the 'Happy', 'Sad', and 'Angry' emotions, as indicated by the high predicted probabilities.

Furthermore, the box in the 'Surprise' row also shows a relatively dark green color. However, there is a slight tendency for it to be predicted as 'Fear'. These observations suggest that the model performs moderately well in recognizing the 'Surprise' emotion. Interestingly, the 'Fear' emotion, apart from being mostly predicted as 'Fear', displays a slight tendency to be predicted as 'Surprise'. This indicates the presence of a potential emotion cluster involving 'Fear' and 'Surprise'. Additionally, the 'Fear' row shows a tendency to be predicted as 'Sad'. These results suggest that predicting 'Fear' and 'Surprise' as standalone emotions may not be entirely reliable. However, when considered as a cluster, they may exhibit stronger probability values. This finding aligns with a study conducted by R. E. Jack et al. [REJS16], which proposed instead of the commonly known six universal facial expression patterns [EF78], which were the emotions tested in the pilot study, four latent expressive patterns: 'Happy' representing positive emotions like happiness and joy, 'Sad' representing negative emotions such as sadness and grief, 'Angry' representing anger and related negative emotions (like disgust and contempt), and 'Surprised/Fearful' combining surprised and fearful expressions.

Regarding the 'Disgust' row, the box suggests that the model primarily predicts 'Disgust', 'Sad', and 'Angry'. The prediction of anger also aligns with the findings of R. E. Jack et al.[REJS16]

Based on the observed results, it appears that the 'enet_b0_8_best_afew.pt' model demonstrates distinct and clear clusters, making it the preferred choice for the facial expression recognition version of the game. These clusters consist of 'Disapproval', 'Distressed/Uncertain', and 'Positive/Approval' emotions.

The 'Disapproval' cluster includes emotions such as 'Anger', 'Disgust', and 'Contempt', which are commonly associated with disapproval. Within the 'Distressed/Uncertain' cluster, we find emotions like 'Sad', 'Fear', and 'Surprise'. 'Sad' and 'Fear' often signify distress, while 'Surprise' typically indicates uncertainty. Notably, 'Fear' and 'Surprise' form a distinct and cohesive cluster, minimizing overlap with other clusters. The 'Positive/Approval' cluster includes 'Happy' and 'Neutral' emotions. 'Happy' represents a positive emotion commonly associated with approval, while 'Neutral' is included because approving something can result in a neutral expression. However, expressions of disapproval, distress, or uncertainty are less likely to manifest as a neutral expression.

The choice of three clusters was made to ensure that each cluster contains at least one emotion that the model recognizes well. This approach increases the likelihood of accurately recognizing the intended emotion or emotional cluster conveyed by the user.

4.2 Main study: comparison level of engagement

This section presents the results of the main study, which are divided into the sections corresponding to the original questionnaires that were combined to create the questionnaire for this study. The results for each individual questionnaire include an analysis of the overall scores and the scores for each specific element within the questionnaire.

4.2.1 System Usability Scale (SUS)

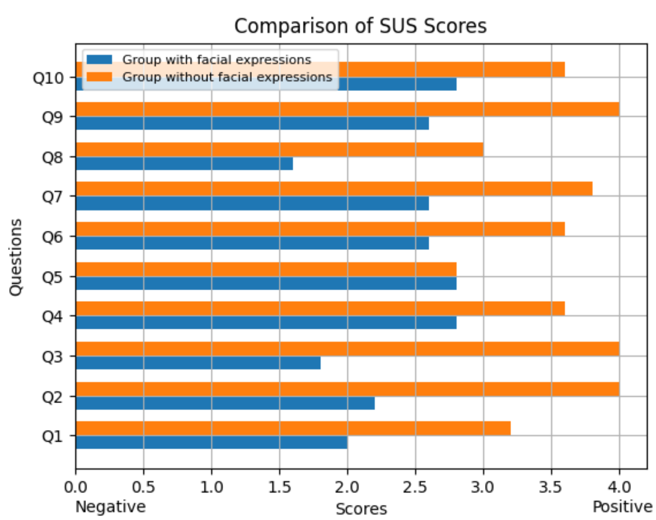

Figure 8: The average scores of the different questions in the SUS, for both versions of the game. The maximum score to be obtained is 4.0, which is very positive.

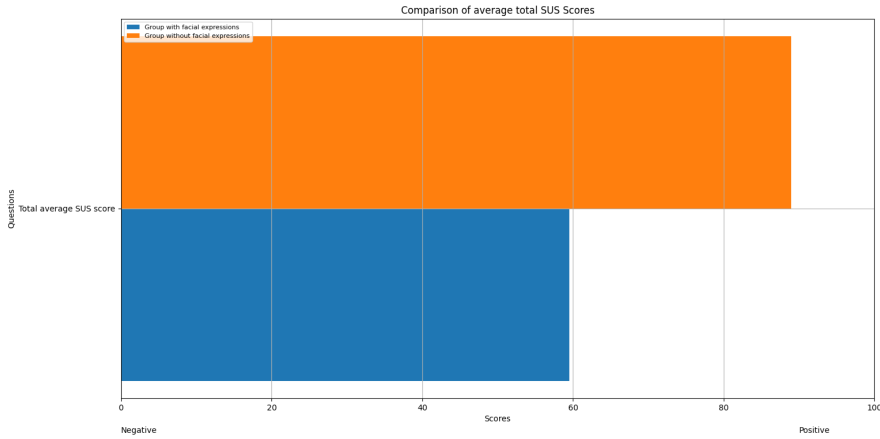

Figure 9: Comparison of average total SUS scores of both versions of the game. The maximum score obtainable is 100, which is very positive.

The results of the System Usability Scale (SUS) part of the questionnaire are presented in Figure 8 and Figure 9. Figure 8 illustrates that the version with facial expression recognition received significantly lower scores on almost all questions related to usability compared to the traditional version. This suggests that the usability of the facial expression recognition version is notably lower. Figure 9 displays a substantial difference in the total SUS scores between the two versions. To further analyze this difference, a paired t-test was conducted. The results of the t-test yielded a p-value of 0.027259654252279066, which is lower than the significance level of 0.05. This indicates that the probability of obtaining the observed results by random chance is less than 5%. Therefore, we can reject the null hypothesis, which assumes no difference in the total SUS scores between the two versions. These findings provide evidence for a statistically significant effect on the SUS scores. In this case, it suggests a negative effect on SUS scores for the version with facial expression recognition. To investigate what causes this difference in usability scores, the question "What do you think made the version with facial expressions less usable?" was added to the follow-up questionnaire.

4.2.2 Game Experience Questionnaire (GEQ)

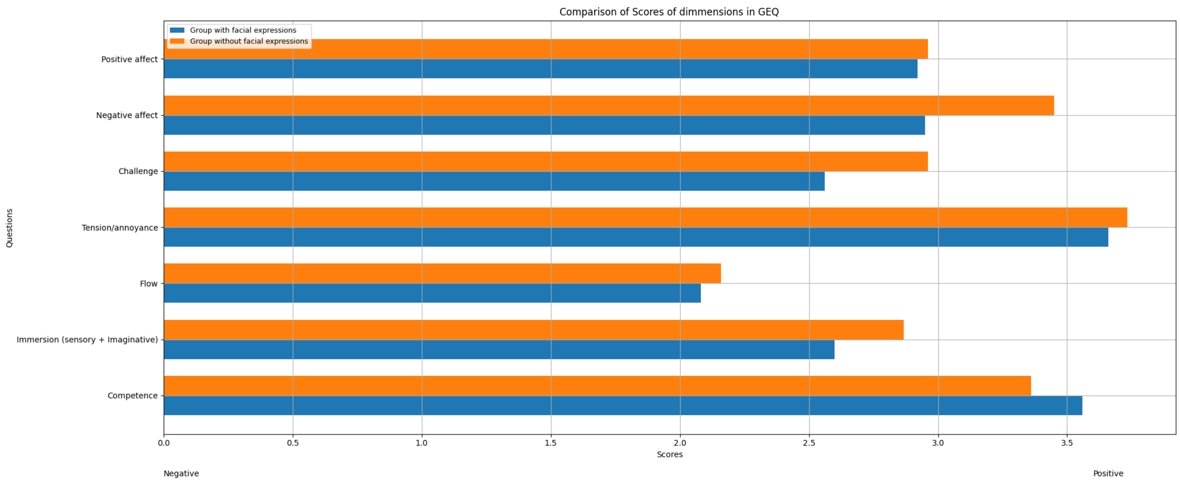

Figure 10: The average scores of the different dimensions of the GEQ, for both versions of the game. The maximum score to be obtained is 4.0, which is very positive.

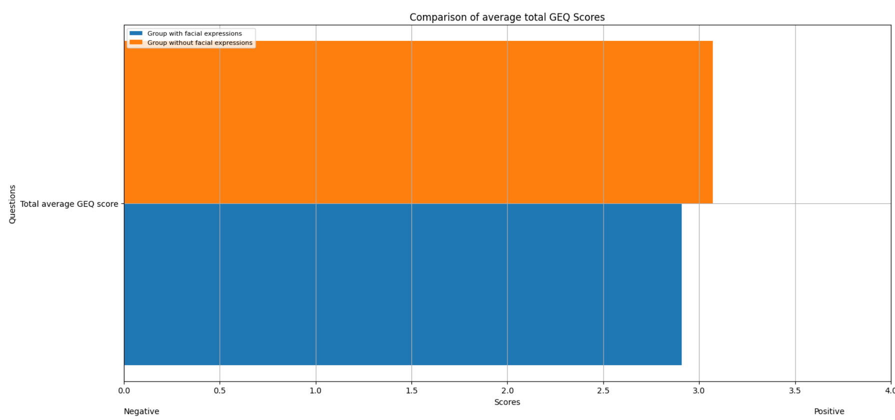

Figure 11: Comparison of average total GEQ scores of both versions of the game. The maximum score obtainable is 4.0, which is very positive.

The results of the Game Experience Questionnaire (GEQ) are presented in Figure 10 and Figure 11. In Figure 10, it is observed that the dimensions of negative affect and challenge show noticeable differences. To further explore these differences, questions related to these dimensions were included in the follow-up questionnaire. These questions inquire about participants' perception of increased challenge and negative feelings in the version with facial expressions. Although there are slight differences across all dimensions, they do not appear to be significant. Figure 11 illustrates that there is not a significant difference in the total scores, suggesting that the measured differences in experience, as captured by the GEQ, are not statistically significant enough to be reported.

4.2.3 PANAS

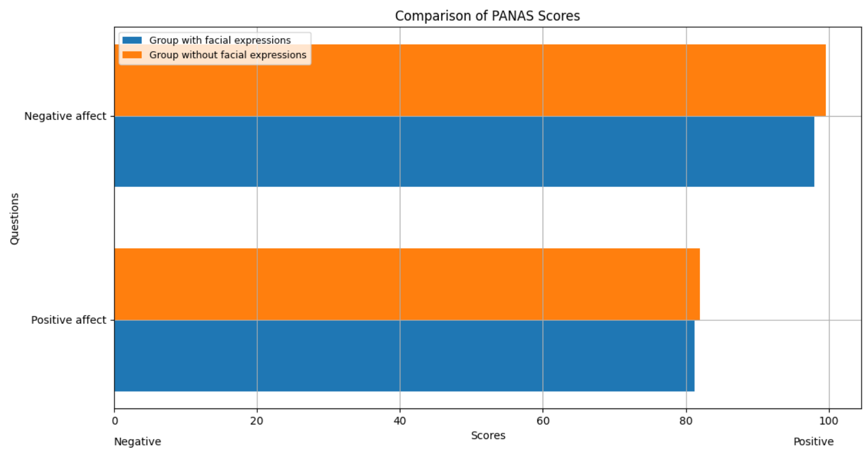

Figure 12: The average scores of negative affect and positive affect measures by PANAS, for both versions of the game. The maximum score to be obtained is 100, which is very positive.

Figure 12 does not indicate any notable differences in both negative affect and positive affect between the two versions. Interestingly, there seems to be a difference in negative affect measured by the GEQ questionnaire. The added question of the follow-up questionnaire about experiencing negative feelings should provide more clarity on this.

4.2.4 NASA-TLX

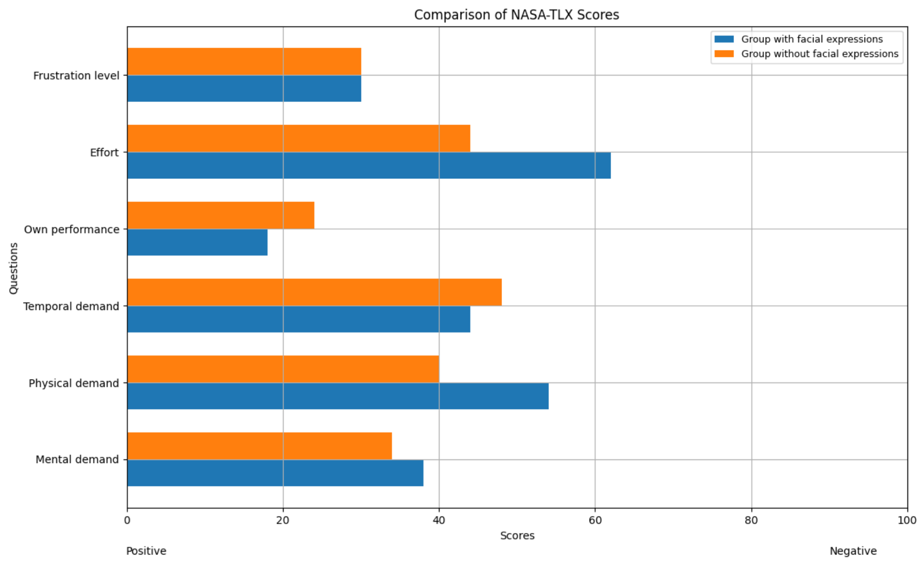

Figure 13: The average scores of the different dimensions of the NASA-TLX, for both versions of the game. The maximum score to be obtained is 100, which is very negative.

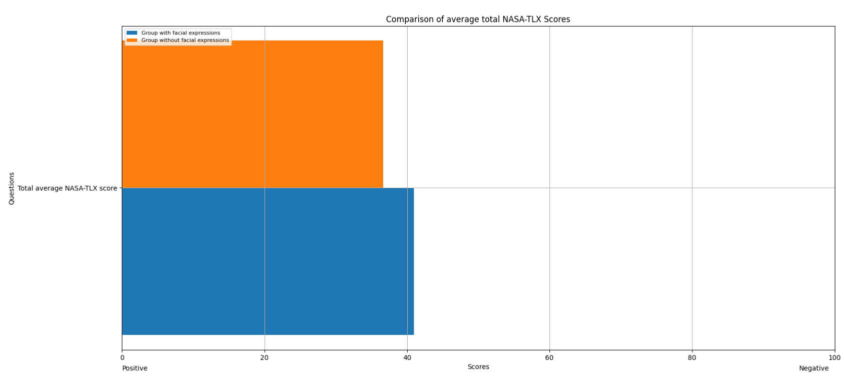

Figure 14: Comparison of average total NASA-TLX scores of both versions of the game. The maximum score obtainable is 100, which is very negative.

The results of the NASA-TLX are presented in Figures 13 and 14. Figure 13 highlights notable visual differences in the amount of effort and physical demand between the two versions of the game. To gain further insights into these differences, the follow-up questionnaire included questions specifically related to these dimensions. Participants were asked, "Do you feel like the version with facial expressions required more effort? If so, why?" and "Do you feel like the version with facial expressions was more physically demanding? If so, why?" Figure 14, however, does not demonstrate a significant difference in the overall average NASA-TLX scores. This suggests that there is no significant difference to report regarding this measure.

4.2.5 follow-up questionnaire

The participants' responses revealed several advantages and disadvantages of using facial expression recognition as a game mechanic. According to the participants, the advantages included the ease of expressing emotions in response to the game and the overall enjoyment. However, some participants did not perceive any advantages to this system.

On the other hand, participants highlighted several disadvantages of using facial expression recognition. They mentioned that their facial expressions could change to a more concentrated expression while reading the questions, potentially leading to the game not accurately detecting the intended emotion. Inconsistency in reading facial expressions and sudden changes in expression were also mentioned as drawbacks. Participants expressed that the traditional version of the game was more practical as it did not require a camera and allowed for the direct selection of emotions. However, they noted that this version was less interactive, interesting, creative, and fun. Participants felt that the usability of the version with facial expression recognition was compromised because facial expressions could change while reading the dialogue options. One participant mentioned that the program sometimes failed to recognize the intended emotions.

Participants agreed that the version with facial expression recognition required more effort and was more physically demanding. They felt that constant attention to their facial expressions was necessary, and at times, they had to put in more effort to convey the desired interaction, resulting in a perceived sense of having to force the facial expressions. Interestingly, although the GEQ results indicated a more negative score on negative affect for the facial expression recognition version, participants did not personally recognize feeling more negative emotions during that version. This suggests that the GEQ might not effectively measure negative affect in this context. The PANAS results indicated no significant difference in positive and negative affect between the two versions of the game.

Regarding suggestions for improving the game, participants proposed various adjustments. One participant suggested preventing frequent changes in dialogue options while reading the questions. This could be done by implementing a locking mechanism. This would allow participants to lock their facial expressions until they are ready to change the dialogue options again. Another participant recommended improving the recognition of different facial expressions. This could be done by using a model with better performance than the one used in this study. Incorporating facial expressions in the non-playable characters (NPCs) to convey emotional states, rather than relying solely on text, was suggested by another participant. Additionally, one participant suggested incorporating speech and audio features to create a more realistic conversation with NPCs. These suggestions could serve as valuable considerations for future research.

Overall, the participants' feedback provided insights into the advantages, disadvantages, and potential improvements of using facial expression recognition as a game mechanic. These findings offer valuable implications for the design and development of similar game systems in the future.

4.3 Discussion

In this section, the findings and implications of the pilot study and the main study will be discussed, as it relates to the research question and hypothesis. The limitations of the research will also be addressed.

4.3.1 Pilot Study: Comparing Facial Expression Recognition Models

The pilot study aimed to compare the performance of three facial expression recognition models: faceAPI, model_filter.h5, and enet_b0_8_best_afew.pt. The findings of the pilot study provided insights into the accuracy and effectiveness of these models in recognizing facial expressions.

The results revealed that the enet_b0_8_best_afew.pt model outperformed the other two models in accurately recognizing facial expressions. The enet_b0_8_best_afew.pt model demonstrated strong recognition capabilities for emotions such as "Happy," "Sad," and "Angry," while also showing moderate success in recognizing "Surprise" and "Fear". However, it struggled to accurately identify and differentiate "Disgusted" emotions.

The enet_b0_8_best_afew.pt model, with its distinct and clear emotion clusters, emerged as the preferred choice for the facial expression recognition version of the game. By incorporating emotions such as "Disapproval," "Distressed/Uncertain," and "Positive/Approval," the model provides a more comprehensive recognition of emotions, enhancing the gameplay experience and usability.

4.3.2 Main Study: Comparison of User Experience

The main study aimed to compare the user experience between the traditional button-based dialogue selection version and the facial expression recognition-based dialogue selection version of the game. The findings shed light on the implications of using facial expression recognition as a game mechanic and its impact on user experience.

The research question was: What is the effect of automatically recognized emotions as a filtering mechanism in a game for the available dialogue options in a conversation with an NPC on user experience? It was centered around the extent to which incorporating facial expression recognition affects user experience in the game.

It was hypothesized that incorporating facial expression recognition as a filtering mechanism for dialogue options would enhance user experience in the game, leading to better player experiences.

Contrary to our hypothesis, the results indicated that the facial expression recognition version of the game received lower scores on usability measures, as assessed by the System Usability Scale (SUS). This suggests that the usability of the facial expression recognition version was notably lower compared to the traditional version. However, it is worth noting that the scores of the other questionnaires did not exhibit a significant difference between the two versions.

These findings indicate that the impact of incorporating facial expression recognition on user experience is more nuanced than anticipated. While the facial expression recognition mechanism offers advantages such as ease of expressing emotions and overall enjoyment, participants also highlighted disadvantages such as changes in facial expressions while reading questions and inconsistency in expression recognition. The traditional version of the game was perceived as more practical but less interactive, interesting, creative, and fun.

Participants may have perceived the version with facial expression recognition as more challenging, physically demanding, and effortful due to the novelty of this mechanic. Since such a system has not really been implemented before, the participants did not have prior experience with a similar setup. There could be a learning curve associated with it, and after some time, these perceived disadvantages might decrease to some degree. To confirm this theory, further studies should be conducted.

4.3.3 limitations

It is important to acknowledge the limitations of this research. First, the sample size was small, consisting of only five participants, which may limit the generalizability of the findings. Additionally, the study primarily targeted adults, and the results may not be directly applicable to children or individuals with different cultural backgrounds. The use of self-report measures introduces subjectivity and potential biases. Furthermore, the facial expression recognition models used in the study may have limitations in accurately capturing subtle nuances of facial expressions. The novelty of the implemented mechanic was also not considered.

5 Conclusion and further research

In conclusion, the pilot study demonstrated that the enet_b0_8_best_afew.pt model performed well in recognizing facial expressions, leading to its selection for the facial expression recognition version of the game. The main study revealed differences in usability between the two game versions, with the traditional version scoring higher on usability scores, as assessed by the System Usability Scale(SUS). However, there were no significant differences in the overall scores of the Game Experience Questionnaire (GEQ), Positive and Negative Affect Scale (PANAS), and NASA Task Load Index (NASA-TLX) between the two versions. The feedback from participants shed light on the advantages and disadvantages of using facial expression recognition as a game mechanic and provided valuable insights for future improvements.

The results suggest that incorporating facial expression recognition as a filtering mechanism offers advantages such as ease of expressing emotions and overall enjoyment. However, challenges related to changes in facial expressions during reading, inconsistency in expression recognition, and other usability issues need to be addressed to optimize the user experience.

Despite the limitations, this research contributes to the understanding of the impact of facial expression recognition on user experience in video games. The findings can inform the design and development of similar game systems, allowing for more engaging and enjoyable gameplay experiences.

This study provides valuable insights for further research in this area. The following list shows the problems and disadvantages of the game version with facial expression recognition that were found during the study and possible solutions to solve these for further research.

- The detected emotion could change while reading the dialogue options, which required the players to constantly pay attention to their facial expressions. Possible solutions could be:
  - Implement a locking mechanism that gives the player the ability to lock the desired emotion when they do not want it to change, and unlock when they want the facial expression recognition to work again. However, this could potentially overcomplicate the mechanic or take the player out of a state of flow.
  - Use the detection of the neutral expression by the model to detect whether the player is reading. It should be examined if players indeed portray a neutral expression while reading.
  - Implement an additional model that could detect gaze. When gaze is detected, the previously detected emotions should not change.
- There appeared to be inconsistencies and sudden changes in the detected facial expressions. The expressions were also not always detected accurately. Possible solutions could be:
  - Utilize a model with better performance than the enet model.
  - Implement different clusters of emotions in the game. There might be clusters that would give more optimal performance.

Certain aspects of the study itself could also be improved. First, expanding the sample size and diversifying the participant pool would enhance the generalisability of the findings. Additionally, considering different age groups, cultural backgrounds, and gaming experiences could provide a more comprehensive understanding of the impact of facial expression recognition on user experience.

Furthermore, conducting longitudinal studies to examine the long-term effects of facial expression recognition in games and assessing its impact on player experience would provide deeper insights into the potential benefits and drawbacks of this game mechanic.

In further research, the proposed improvements of the game by the participants could also be examined. For example, investigating the impact of incorporating speech and audio features alongside facial expressions in the game could enhance the realism and engagement of the player-NPC interactions. Additionally, incorporating facial expressions in NPCs as a reaction to the player's choices can lead to more immersive game experiences.

References

[AVSM22] L. V. Savchenko A. V. Savchenko and I. Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 2022.

[Bra] Google Brain. Tensorflow. retrieved from https://www.tensorflow.org/.

[Bro95] J. Brooke. Sus - a quick and dirty usability scale. 1995.

[BV17] B. Bontchev and D. Vassileva. Affect-based adaptation of an applied video game for educational purposes. *Interactive Technology and Smart Education*, 14:31–49, 2017.

[Coma] Numpy Community. Numpy. retrieved from https://numpy.org/install/.

[comb] Pandas community. Pandas. retrieved from https://pandas.pydata.org/docs/.

[Coo98] Graham Cooper. Research into cognitive load theory and instructional design at unsw. 1998.

[Csi90] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. 1990.

[Dev] Matplotlib Developers. Matplotlib. retrieved from https://matplotlib.org.

[DH00] Y. Douglas and A. Hargadon. The pleasure principle: immersion, engagement, flow. *Proceedings of the eleventh ACM on Hypertext and hypermedia*, pages 153–160, 2000.

[Dol] C. Dolphin. Flask-cors: A flask extension for handling cross origin resource sharing (cors). retrieved from https://flask-cors.readthedocs.io/.

[DS16] M. Dubey and Prof. L. Singh. Automatic emotion recognition using facial expression: A review. *International Research Journal of Engineering and Technology (IRJET)*, 3:488–492, 2016.

[DWT88] L. A. Clark D. Watson and A. Tellegen. Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, 54:1063–1070, 1988.

[EF78] P. Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial movement. 1978.

[Esl] T. Esler. facenet_pytorch [model: Mtcnn]. retrieved from https://pypi.org/project/facenet-pytorch/.

[Foua] Python Software Foundation. Base64. retrieved from https://docs.python.org/3/library/base64.html.

[Foub] Python Software Foundation. json. retrieved from https://docs.python.org/3/library/json.html.

[fS18a] International Organization for Standardization. Ergonomics of human-system interaction — part 11: Usability: Definitions and concepts (iso 9241-11:2018). 2018. section 3.2.3, retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en".

[fS18b] International Organization for Standardization. Ergonomics of human-system interaction — part 11: Usability: Definitions and concepts (iso 9241-11:2018). 2018. section 3.1.1, retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en".

[FSP15] D. Kalenichenko F. Schroff and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv:1503.03832*, 2015.

[GCP06] D. Grandjean G. Chanel, J. Kronegg and T. Pun. Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. *Proceedings of the International Workshop on Multimedia Content Representation, Classification And Security*, 4105:530–537, 2006.

[GCP09] M. Soleymani G. Chanel, J.J.M Kierkels and T. Pun. Short-term emotion assessment in a recall paradigm. *Int. J. Hum. Comput. Stud.*, 67:607–627, 2009.

[GCP20] D. Grandjean G. Chanel, J. Kronegg and T. Pun. Self-supervised learning for ecg-based emotion recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3217–3221, 2020.

[Has08] M. Hassenzahl. User experience (ux): Towards an experiential perspective on product quality. *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine on - IHM.*, 2008.

[HS88] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139–183, 1988.

[IC12] A. B. Ingale and D. S. Chaudhari. Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, 2:235–238, 2012.

[JHB09] K. A. Curtiss E. McBroom K. M. Burkhart J. N. Pidruzny J. H. Brockmyer, C. M. Fox. The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45:624–634, 2009.

[KGA05] A. Dix K. Gilleade and J. Allanson. Affective videogames and modes of affective gaming: Assist me, challenge me, emote me. *DiGRA 2005: Changing Views*, 2005.

[KRM10] M. Ohkura K. Rattanyu and M. Mizukawa. Emotion monitoring from physiological signals for service robots in the living space. *Proceedings of the ICCAS*, pages 580—583, 2010.

[KZQ16] Z. Li K. Zhang, Z. Zhang and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016.

[Lun] F. Lundh. Pillow (pil fork). retrieved from https://python-pillow.org/.

[MGR20] D. Maggiorini M. Granato, D. Gadia and L. A. Ripamonti. An empirical study of players' emotions in vr racing games based on a dataset of physiological data. *Multimed Tools Appl*, 79, 2020.

[Mü] V. Mühler. face-api.js. retrieved from https://github.com/justadudewhohacks/face-api.js.

[Ope] OpenCV. Opencv: Open source computer vision library. retrieved from https://opencv.org/.

[Pal] Pallets. Flask: A python microframework. retrieved from https://flask.palletsprojects.com/.

[REJS16] I. Delis O. G. B. Garrod R. E. Jack, W. Sun and P. G. Schyns. Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General*, 145:708–730, 2016.

[San] D. Sandberg. facenet. *Github*. retrieved from https://github.com/davidsandberg/facenet.

[Sav] A. V. Savchenko. hsemotion.facial_emotions [model: enet_b0_8_best_afew in hsemotion-recognizer]. retrieved from https://github.com/HSE-asavchenko/hsemotion.

[Sav21] A. V. Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *Proceedings of the 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE, 2021.

[Sav22] A. V. Savchenko. Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2359–2366, June 2022.

[Sav23] A. V. Savchenko. MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data. In *Proceedings of the European Conference on Computer Vision (ECCV 2022) Workshops*, pages 45–59. Springer, 2023.

[Sha] A. Shawon. model_filter.h5. retrieved from https://www.kaggle.com/code/shawon10/facial-expression-detection-cnn/output.

[SPC14] F. Odone A. Verri S. Piana, A. Staglianò and A. Camurri. Real-time automatic emotion recognition from body gestures. *ArXiv*, 2014.

[TDN19] R. Ragel T. Dissanayake, Y. Rajapaksha and I. Nawinne. An ensemble learning approach for electrocardiogram sensor based human emotion recognition. *Sensors*, 19, 2019.

[VL07] O. Villon and C. Lisetti. Toward recognizing individual's subjective emotion from physiological signals in practical application. *Proceedings of the Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*, pages 357–362, 2007.

[VVN14] K. i. Fukui V. Vachiratamporn, K. Moriyama and M. Numao. An implementation of affective adaptation in survival horror games. *IEEE Conference on Computational Intelligence and Games*, pages 1–8, 2014.

[Was] M. Waskom. Seaborn. retrieved from https://seaborn.pydata.org/.

[YIZ22] S. Otmane A. Khababa M. Masmoudi Y. Izountar, S. Benbelkacem and N. Zenati. Vr-peer: A personalized exer-game platform based on emotion recognition. *Electronics*, 11, 2022.

[ZB20] S. Zhou and H. Beigi. A transfer learning method for speech emotion recognition from automatic speech recognition. *ArXiv*, 2020.

A Usage of chatGPT

In this bachelor thesis, chatGPT was used for multiple causes. The primary use of chatGPT was to rephrase/rewrite texts written by myself. This was usually done by using the following prompt: "Could you rewrite the following section as if it were in an [information about section] for a bachelor's thesis. Please do not make the sentences overly complex, keep it concrete and try not to use too many complex words: [text]". The returned text was then copied and checked for odd words. Furthermore, chatGPT was used to ask for a recommended structure in particular sections, and to validate whether a particular section was well written using the following command: "I have written a [section] for my bachelor thesis. Do you think the following [section] could be considered complete and well-written enough?: [text]". ChatGPT was also used to bug-fix code at times. It was also used to ask for an implementation method of the API. It then suggested using Flask and Flask_Cors. Furthermore, I have occasionally asked chatGPT for references to papers about particular topics, however, chatGPT mostly gave faulty information. It was also used to get inspiration for the game. I asked it for minigame ideas, possible names for the characters and jobs. ChatGPT was also used to create the dialogue for the game. The exact prompts to do this will be noted in different appendices.

B chatGPT prompt to create persona

The prompt used to create a persona for a fictional character in the game was as follows:

"Give a persona (character) description according to the following characteristics: "A [young/adult/elder] [male/female] with an [closed/open], [conscientious/sloppy], [extravert/introvert], [argumentative/agreeable], [neurotic/calm] personality (according to OCEAN personality theory), living in a prehistoric tribe on a tropical island, with the profession of [job], and a family setting with [amount of kids and presence of wife/husband]""

C chatGPT prompt to create dialogue options

The prompt used to create six different dialogue options within different emotion classes, two dialogue options per emotion cluster, was as follows:

[Situation]. [Main character] in her/his turn always has six options to choose from: two disapproving options [optional information as to what is the case in those options], two distressed options [optional information as to what is the case in those options], and two positive approving options[optional information as to what is the case in those options]. Take into account their persona very carefully in the dialogue. Give the six options for [Main character]'s turn. Please try to keep the options short. They should not be longer than 1/2 sentences.

D chatGPT prompt to create responses to dialogue options

The characters in the games sometimes respond differently to different dialogue options selected by the player. To create these responses, the following prompt was used:

”For each option, [Character] has to give a response. Please generate a response that reacts to [Main character] their response and make sure [Character] [provide information about something the character has to include in their responses]. Take into account [Character] their persona very carefully. To be clear, there should be 6 responses in total, one to each response from [Main character]. Don’t make the responses too long.”

E Implementation details for API

In order to develop the API, it was necessary to consider the facial expression recognition process. Typically, facial expression recognition models analyze emotions based on individual frames. Thus, the API needed to accept an image as input. Additionally, most models lack the capability to identify the location of the face within the frame. Consequently, a face detection model was also required to identify faces within the frame. As for the output, these models usually generate JSON format results containing emotion probabilities. Therefore, the API needed to provide JSON format output, including probabilities for each recognized emotion.

Each model within the API required a unique ”address” for access. When invoking the API, the user must provide their image which the API will receive in the form of a base64-encoded string. Within the API, several checks are performed on the user’s input. First, it verifies whether the input is provided under the correct parameter name and ensures that it is not an empty string. If a string is found under the correct parameter name, the API proceeds to validate whether it is a valid base64 string. Once validated, the string is decoded into grayscale, and the resulting image is passed to the face detection model. This model is capable of detecting multiple faces within a frame, but the API is designed to return emotion probabilities for only a single face. Therefore, only the first detected face image is utilized. The image of the detected face is then converted to a tensor of the right size for that particular model, and passed to the facial expression model. Subsequently, the model generates a JSON format containing the probabilities for different emotions. Finally, the API ensures the correct structure of the output before returning the emotion probabilities to the user.

The API implementation was carried out in Python. The Flask framework [Pal], along with the Flask-Cors package [Dol], was utilized to handle API requests. Flask-Cors was specifically employed to address any potential issues related to the same origin policy. The base64 [Foua] package was utilized to decode the base64 string received as input. Image processing tasks were performed using the cv2 [Ope] and PIL [Lun] packages. Additionally, numpy [Coma] and TensorFlow [Bra] libraries were utilized to transform the images into arrays and tensors. TensorFlow was also employed to load the ’model_filter.h5’ model for further processing within the API.

The API was manually tested to evaluate its functionality. This involved using CURL statements to send requests containing images with known facial expressions. The probabilities returned by the API were inspected and compared against the known facial expressions present in the images. This manual validation was performed to assess if the facial expression recognition process was implemented in the right way. Subsequently, the performance of the models themselves was further assessed in a pilot study which will be discussed in later sections.

The API was not deployed on a public network, therefore, limited emphasis was placed on implementing extensive security measures within the API. When the API was utilized, the API operated on a desktop located within the same network as the program utilizing it.

Due to privacy considerations, the images sent to the API were not stored in a database. Consequently, the images were processed in real-time, ensuring that only the API had access to the frames containing users' faces. The API was not publicly deployed to further safeguard privacy. This decision was made to prevent unauthorized individuals from intercepting or eavesdropping on the API requests.

F Implementation details for Pilot Study

Within the code of the main page, the first step involved initializing a canvas to store the frame to be processed. Subsequently, the faceAPI models were loaded. If all models were successfully loaded, the video stream from the webcam was started.

Once the video stream was active, facial expression probabilities returned by the faceAPI framework were requested every 500 milliseconds. This number was chosen by calculating how long processing a frame by three different models approximately takes. If the researcher pressed the submit button for a specific emotion and if the faceAPI successfully detected the participant's face, the frame was saved in the canvas. It was then converted to a base64 string and sent to both models of the API.

The probabilities returned by all three models were saved in separate JSON files, named according to the respective model. For each emotion, when the submit button was pressed, the participant's face frame was sent 10 times. The decision to save probabilities for 10 frames instead of a single frame was made to enhance reliability. Capturing only one frame increases the risk of capturing a moment when the participant's face may have been in motion, potentially leading to inaccurate probabilities reflecting unintended emotions. By averaging the emotion probabilities over 10 frames, the likelihood of not capturing the emotion the participant wants to convey due to the participant's face being in motion is minimized, thus improving the reliability of the collected data.

After the 10 frames are analysed, the displayed emotion name changed, and the new name was highlighted in red to enhance visibility. The emotion name in the JSON file was also updated accordingly.

Once the probabilities of the 10 frames for all six emotions were stored in the respective JSON files, the three JSON files were automatically downloaded. Subsequently, the participant was redirected to a thank you page.

G SUS

System Usability Scale

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5. I found the various functions in this system were well integrated	1	2	3	4	5
6. I thought there was too much inconsistency in this system	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8. I found the system very cumbersome to use	1	2	3	4	5
9. I felt very confident using the system	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

Using SUS

The SU scale is generally used after the respondent has had an opportunity to use the system being evaluated, but before any debriefing or discussion takes place. Respondents should be asked to record their immediate response to each item, rather than thinking about items for a long time.

All items should be checked. If a respondent feels that they cannot respond to a particular item, they should mark the centre point of the scale.

Scoring SUS

SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own.

To calculate the SUS score, first sum the score contributions from each item. Each item's score contribution will range from 0 to 4. For items 1,3,5,7,and 9 the score contribution is the scale position minus 1. For items 2,4,6,8 and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall value of SU.

SUS scores have a range of 0 to 100.

The following section gives an example of a scored SU scale.

H GEQ

2. Game Experience Questionnaire – Core Module

Please indicate how you felt while playing the game for each of the items,
on the following scale:

not at all	slightly	moderately	fairly	extremely
0	1	2	3	4
< >	< >	< >	< >	< >

1. I felt content
2. I felt skilful
3. I was interested in the game's story
4. I thought it was fun
5. I was fully occupied with the game
6. I felt happy
7. It gave me a bad mood
8. I thought about other things
9. I found it tiresome
10. I felt competent
11. I thought it was hard
12. It was aesthetically pleasing
13. I forgot everything around me
14. I felt good
15. I was good at it
16. I felt bored
17. I felt successful
18. I felt imaginative
19. I felt that I could explore things
20. I enjoyed it
21. I was fast at reaching the game's targets
22. I felt annoyed
23. I felt pressured
24. I felt irritable
25. I lost track of time
26. I felt challenged
27. I found it impressive
28. I was deeply concentrated in the game
29. I felt frustrated
30. It felt like a rich experience
31. I lost connection with the outside world
32. I felt time pressure

33 I had to put a lot of effort into it

6. Scoring guidelines

Scoring guidelines GEQ Core Module

The Core GEQ Module consists of seven components; the items for each are listed below.

Component scores are computed as the average value of its items.

Competence: Items 2, 10, 15, 17, and 21.

Sensory and Imaginative Immersion: Items 3, 12, 18, 19, 27, and 30.

Flow: Items 5, 13, 25, 28, and 31.

Tension/Annoyance: Items 22, 24, and 29.

Challenge: Items 11, 23, 26, 32, and 33.

Negative affect: Items 7, 8, 9, and 16.

Positive affect: Items 1, 4, 6, 14, and 20.

Scoring guidelines GEQ In-Game version

The In-game Module consists of seven components, identical to the core Module. However, only two items are used for every component. The items for each are listed below.

Component scores are computed as the average value of its items.

Competence: Items 2 and 9.

Sensory and Imaginative Immersion: Items 1 and 4.

Flow: Items 5 and 10.

Tension: Items 6 and 8.

Challenge: Items 12 and 13.

Negative affect: Items 3 and 7.

Positive affect: Items 11 and 14.

Scoring guidelines GEQ Social Presence Module

The Social Presence Module consists of three components; the items for each are listed below.

Component scores are computed as the average value of its items.

Psychological Involvement – Empathy: Items 1, 4, 8, 9, 10, and 13.

Psychological Involvement – Negative Feelings: Items 7, 11, 12, 16, and 17.

Behavioural Involvement: Items 2, 3, 5, 6, 14, and 15.

Scoring guidelines GEQ Post-game Module

The post-game Module consists of four components; the items for each are listed below.

Component scores are computed as the average value of its items.

Positive Experience: Items 1, 5, 7, 8, 12, 16.

Negative experience: Items 2, 4, 6, 11, 14, 15.

Tiredness: Items 10, 13.

Returning to Reality: Items 3, 9, and 17.

I PANAS

"Blue Monday" phenomenon. *Journal of Personality and Social Psychology, 49,* 129–134.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma & J. D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 681–706). Hillsdale, NJ: Erlbaum.

Tessler, R., & Mechanic, D. (1978). Psychological distress and perceived health status. *Journal of Health and Social Behavior, 19,* 254–262.

Warr, P. (1978). A study of psychological well-being. *British Journal of Psychology, 69,* 111–121.

Warr, P., Barter, J., & Brownbridge, G. (1983). On the independence of positive and negative affect. *Journal of Personality and Social Psychology, 44,* 644–651.

Watson, D. (in press). The vicissitudes of mood measurement: Effects of varying descriptors, time frames, and response formats on measures of Positive and Negative Affect. *Journal of Personality and Social Psychology.*

Watson, D. (1988). Intraindividual and interindividual analyses of Positive and Negative Affect: Their relation to health complaints, perceived stress, and daily activities. *Journal of Personality and Social Psychology, 54,* 1020–1030.

Watson, D., & Clark, L. A. (1984). Negative Affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin, 96,* 465–490.

Watson, D., Clark, L. A., & Carey, G. (in press). Positive and Negative Affectivity and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology.*

Watson, D., Clark, L. A., & Tellegen, A. (1984). Cross-cultural convergence in the structure of mood: A Japanese replication and a comparison with U.S. findings. *Journal of Personality and Social Psychology, 47,* 127–144.

Watson, D., & Pennebaker, J. W. (in press). Health complaints, stress, and distress: Exploring the central role of negative affectivity. *Psychological Review.*

Watson, D., & Tellegen, A. (1985) Toward a consensual structure of mood. *Psychological Bulletin, 98,* 219–235.

Wills, T. A. (1986). Stress and coping in early adolescence: Relationships to substance use in urban school samples. *Health Psychology, 5,* 503–529.

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology, 43,* 111–122.

Appendix

The PANAS

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent [INSERT APPROPRIATE TIME INSTRUCTIONS HERE]. Use the following scale to record your answers.

1	2	3	4	5
very slightly or not at all	a little	moderately	quite a bit	extremely

_____ interested	_____ irritable
_____ distressed	_____ alert
_____ excited	_____ ashamed
_____ upset	_____ inspired
_____ strong	_____ nervous
_____ guilty	_____ determined
_____ scared	_____ attentive
_____ hostile	_____ jittery
_____ enthusiastic	_____ active
_____ proud	_____ afraid

We have used PANAS with the following time instructions:

Moment	(you feel this way right now, that is, at the present moment)
Today	(you have felt this way today)
Past few days	(you have felt this way during the past few days)
Week	(you have felt this way during the past week)
Past few weeks	(you have felt this way during the past few weeks)
Year	(you have felt this way during the past year)
General	(you generally feel this way, that is, how you feel on the average)

Received May 10, 1987
Revision received September 14, 1987
Accepted November 11, 1987 ■

J NASA-TLX

APPENDIX A: Sample Application of the NASA-TLX.

EXAMPLE:

COMPARE WORKLOAD OF TWO TASKS THAT REQUIRE A SERIES OF DISCRETE RESPONSES. THE PRIMARY DIFFICULTY MANIPULATION IS THE INTER-STIMULUS INTERVAL (ISI) – (TASK 1 = 500 msec. TASK 2 = 300 msec)

PAIR-WISE COMPARISONS OF FACTORS:

INSTRUCTIONS: SELECT THE MEMBER OF EACH PAIR THAT PROVIDED THE MOST SIGNIFICANT SOURCE OF WORKLOAD VARIATION IN THESE TASKS

PD / MD	TD / PD	TD / FR
TD / MD	OP / PD	TD / EF
OP / MD	FR / PD	OP / FR
FR / MD	EF / PD	OP / EF
EF / MD	TD / OP	EF / FR

TALLY OF IMPORTANCE SELECTIONS

MD III	=	3
PD	=	0
TD IIIII	=	5
OP I	=	1
FR III	=	3
EF III	=	3
SUM	=	15

RATING SCALES:

INSTRUCTIONS: PLACE A MARK ON EACH SCALE THAT REPRESENTS THE MAGNITUDE OF EACH FACTOR IN THE TASK YOU JUST PERFORMED

DEMANDS	RATINGS FOR TASK 1:			RATING		WEIGHT		PRODUCT
MD	LOW	I__x__________I	HIGH	30	X	3	=	90
PD	LOW	I_x___________I	HIGH	15	X	0	=	0
TD	LOW	I_________x___I	HIGH	60	X	5	=	150
OP	EXCL	I____x________I	POOR	40	X	1	=	40
FR	LOW	I__x__________I	HIGH	30	X	3	=	90
EF	LOW	I_____x_______I	HIGH	40	X	3	=	120
				SUM			=	490
				WEIGHTS (TOTAL)			=	15
				MEAN WWL SCORE			=	32

DEMANDS	RATINGS FOR TASK 2:			RATING		WEIGHT		PRODUCT
MD	LOW	I__x__________I	HIGH	30	X	3	=	90
PD	LOW	I__x__________I	HIGH	25	X	0	=	0
TD	LOW	I___________x_I	HIGH	70	X	5	=	350
OP	EXCL	I______x______I	POOR	50	X	1	=	50
FR	LOW	I______x______I	HIGH	50	X	3	=	150
EF	LOW	I__x__________I	HIGH	30	X	3	=	90
				SUM			=	730
				WEIGHTS (TOTAL)			=	15
				MEAN WWL SCORE			=	49

RESULTS:

SUBSCALES PINPOINT SPECIFIC SOURCE OF WORKLOAD VARIATION BETWEEN TASKS (TD). THE WWL SCORE REFLECTS THE IMPORTANCE OF THIS AND OTHER FACTORS AS WORKLOAD-DRIVERS AND THEIR SUBJECTIVE MAGNITUDE IN EACH TASK

K Questionnaire for study

Usability and engagement questionnaire

This questionnaire examines the level of engagement and the usability of two versions of a dialogue based game. In one version, facial expressions were used as a filter mechanism in a (mostly) dialogue based game. The other version utilized the old fashioned way of filtering, namely with buttons. This questionnaire will be used to study the effect of using facial expressions as a filter mechanism on player engagement and the usability of a dialogue based game.

* Verplichte vraag

1. What is your gender? *

   Markeer slechts één ovaal.

   - Male
   - Female
   - Prefer not to say
   - Anders: ______________________

2. How old are you? *

   ____

Game **with** Facial expressions

This part of the questionaire will test your level of engagement and the usability of the version of the game **with** facial expression filtering.

Usability (SUS)

This part of the questionnaire measures how usable the UI was for you.

3. I think that I would like to use this system frequently *

   Markeer slechts één ovaal.

   Strongly disagree

   1 ◯
   2 ◯
   3 ◯
   4 ◯
   5 ◯

   Strongly agree

4. I found the system unnecessarily complex. *

   Markeer slechts één ovaal.

   Strongly disagree

   1 ◯
   2 ◯
   3 ◯
   4 ◯
   5 ◯

   Strongly agree

5. I thought the system was easy to use. *

Markeer slechts één ovaal.

	Strongly disagree
1	◯
2	◯
3	◯
4	◯
5	◯
	Strongly agree

6. I think that I would need the support of a technical person to be able to use this system. *

Markeer slechts één ovaal.

	Strongly disagree
1	◯
2	◯
3	◯
4	◯
5	◯
	Strongly agree

7. I found the various functions in this system were well integrated. *

Markeer slechts één ovaal.

	Strongly disagree
1	◯
2	◯
3	◯
4	◯
5	◯
	Strongly agree

8. I thought there was too much inconsistency in this system. *

Markeer slechts één ovaal.

	Strongly disagree
1	◯
2	◯
3	◯
4	◯
5	◯
	Strongly agree

9. I would imagine that most people would learn to use this system very quickly. *

Markeer slechts één ovaal.

Strongly disagree

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Strongly agree

10. I found the system very cumbersome to use. *

Markeer slechts één ovaal.

Strongly disagree

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Strongly agree

11. I felt very confident using the system. *

Markeer slechts één ovaal.

Strongly disagree

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Strongly agree

12. I needed to learn a lot of things before I could get going with this system. *

Markeer slechts één ovaal.

Strongly disagree

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Strongly agree

13. I felt content *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

14. I felt skilful *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

15. I was interested in the game's story *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

16. I thought it was fun *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

17. I was fully occupied with the game *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

18. I felt happy *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

19. It gave me a bad mood *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

20. I thought about other things *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

21. I found it tiresome *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

22. I felt competent *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

23. I thought it was hard *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

24. It was aesthetically pleasing *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

25. I forgot everything around me *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

26. I felt good *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

27. I was good at it *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

extremely

28. I felt bored *

Markeer slechts één ovaal.

Not at all

1

2

3

4

5

extremely

29. I felt succesful *

Markeer slechts één ovaal.

Not at all

1

2

3

4

5

extremely

30. I felt imaginative *

Markeer slechts één ovaal.

Not at all

1

2

3

4

5

extremely

31. I felt that I could explore things *

Markeer slechts één ovaal.

Not at all

1

2

3

4

5

extremely

32. I enjoyed it *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

33. I was fast at reaching the game's targets *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

34. I felt annoyed *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

35. I felt pressured *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

36. I felt irritable *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

37. I lost track of time *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

38. I felt challenged *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

39. I found it impressive *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

40. I was deeply concentrated in the game *

Markeer slechts één ovaal.

Not at all

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

extremely

41. I felt frustrated *

Markeer slechts één ovaal.

Not at all

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

extremely

42. I felt like a rich experience *

Markeer slechts één ovaal.

Not at all

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

extremely

43. I lost connection with the outside world *

Markeer slechts één ovaal.

Not at all

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

extremely

44. I felt time pressure *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

45. I had to put a lot of effort into it *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

Positive Affect subscale (PANAS-P)
Rate how much you have experienced the following emotions while playing the game.

46. Interested *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
Extremely

47. Excited *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

48. Strong *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

49. Enthusiastic *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

50. Proud *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

51. Alert *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

52. Inspired *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

53. Determined *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

54. Attentive *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

55. Active *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Extremely

Negative Affect subscale (PANAS-N)
Rate how much you have experienced the following emotions while playing the game.

56. Distressed *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Extremely

57. Upset *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Extremely

58. Guilty *

Markeer slechts één ovaal.

Not at all	
1	◯
2	◯
3	◯
4	◯
5	◯
Extremely	

59. Scared *

Markeer slechts één ovaal.

Not at all	
1	◯
2	◯
3	◯
4	◯
5	◯
Extremely	

60. Hostile *

Markeer slechts één ovaal.

Not at all	
1	◯
2	◯
3	◯
4	◯
5	◯
Extremely	

61. Irritable *

Markeer slechts één ovaal.

Not at all	
1	◯
2	◯
3	◯
4	◯
5	◯
Extremely	

62. Ashamed *

Markeer slechts één ovaal.

Not at all

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

Extremely

63. Nervous *

Markeer slechts één ovaal.

Not at all

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

Extremely

64. Jittery *

Markeer slechts één ovaal.

Not at all

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

Extremely

65. Afraid *

Markeer slechts één ovaal.

Not at all

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

Extremely

The NASA Task Load Index (NASA-TLX)
This part of the questionnaire measures the cognitive load of the game.

66. How mentally demanding was the game? *

Markeer slechts één ovaal.

Very low
1 ⬭
2 ⬭
3 ⬭
4 ⬭
5 ⬭
6 ⬭
7 ⬭
8 ⬭
9 ⬭
10 ⬭
Very high

67. How physically demanding was the game? *

Markeer slechts één ovaal.

Very low
1 ⬭
2 ⬭
3 ⬭
4 ⬭
5 ⬭
6 ⬭
7 ⬭
8 ⬭
9 ⬭
10 ⬭
Very high

68. How hurried or rushed was the pace of the game? *

Markeer slechts één ovaal.

Very low
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
6 ◯
7 ◯
8 ◯
9 ◯
10 ◯
Very high

69. How successful do you think you were in accomplishing the goals of the game? *

Markeer slechts één ovaal.

Very low
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
6 ◯
7 ◯
8 ◯
9 ◯
10 ◯
Very high

70. How hard did you have to work to accomplish your level of performance? *

Markeer slechts één ovaal.

Very low

1

2

3

4

5

6

7

8

9

10

Very high

71. How insecure, discouraged, irritated, stressed, and annoyed were you during the game? *

Markeer slechts één ovaal.

Very low

1

2

3

4

5

6

7

8

9

10

Very high

Game **without** Facial expressions

This part of the questionaire will test your level of engagement and the usability of the version of the game **without** facial expression filtering.

Usability (SUS)

This part of the questionnaire measures how usable the UI was for you.

72. I think that I would like to use this system frequently *

Markeer slechts één ovaal.

Strongly disagree

1
2
3
4
5

Strongly agree

73. I found the system unnecessarily complex. *

Markeer slechts één ovaal.

Strongly disagree

1
2
3
4
5

Strongly agree

74. I thought the system was easy to use. *

Markeer slechts één ovaal.

Strongly disagree

1
2
3
4
5

Strongly agree

75. I think that I would need the support of a technical person to be able to use this system. *

Markeer slechts één ovaal.

Strongly disagree

1
2
3
4
5

Strongly agree

76. I found the various functions in this system were well integrated. *

Markeer slechts één ovaal.

	Strongly disagree
1	◯
2	◯
3	◯
4	◯
5	◯
	Strongly agree

77. I thought there was too much inconsistency in this system. *

Markeer slechts één ovaal.

	Strongly disagree
1	◯
2	◯
3	◯
4	◯
5	◯
	Strongly agree

78. I would imagine that most people would learn to use this system very quickly. *

Markeer slechts één ovaal.

	Strongly disagree
1	◯
2	◯
3	◯
4	◯
5	◯
	Strongly agree

79. I found the system very cumbersome to use. *

Markeer slechts één ovaal.

	Strongly disagree
1	◯
2	◯
3	◯
4	◯
5	◯
	Strongly agree

80. I felt very confident using the system. *

Markeer slechts één ovaal.

Strongly disagree

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

Strongly agree

81. I needed to learn a lot of things before I could get going with this system. *

Markeer slechts één ovaal.

Strongly disagree

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

Strongly agree

Game experience questionnaire (GEQ)

82. I felt content *

Markeer slechts één ovaal.

Not at all

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

extremely

83. I felt skilful *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

84. I was interested in the game's story *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

85. I thought it was fun *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

86. I was fully occupied with the game *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

87. I felt happy *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

88. It gave me a bad mood *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

89. I thought about other things *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

90. I found it tiresome *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

91. I felt competent *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

92. I thought it was hard *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

93. It was aesthetically pleasing *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

94. I forgot everything around me *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

95. I felt good *

Markeer slechts één ovaal.

	Not at all
1	◯
2	◯
3	◯
4	◯
5	◯
	extremely

96. I was good at it *

Markeer slechts één ovaal.

	Not at all
1	◯
2	◯
3	◯
4	◯
5	◯
	extremely

97. I felt bored *

Markeer slechts één ovaal.

	Not at all
1	◯
2	◯
3	◯
4	◯
5	◯
	extremely

98. I felt succesful *

Markeer slechts één ovaal.

	Not at all
1	◯
2	◯
3	◯
4	◯
5	◯
	extremely

99. I felt imaginative *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

100. I felt that I could explore things *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

101. I enjoyed it *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

102. I was fast at reaching the game's targets *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

103. I felt annoyed *

Markeer slechts één ovaal.

	Not at all
1	
2	
3	
4	
5	
	extremely

104. I felt pressured *

Markeer slechts één ovaal.

	Not at all
1	
2	
3	
4	
5	
	extremely

105. I felt irritable *

Markeer slechts één ovaal.

	Not at all
1	
2	
3	
4	
5	
	extremely

106. I lost track of time *

Markeer slechts één ovaal.

	Not at all
1	
2	
3	
4	
5	
	extremely

107. I felt challenged *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

108. I found it impressive *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

109. I was deeply concentrated in the game *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

110. I felt frustrated *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
extremely

111. I felt like a rich experience *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

112. I lost connection with the outside world *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

113. I felt time pressure *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

114. I had to put a lot of effort into it *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
extremely

Positive Affect subscale (PANAS-P)
Rate how much you have experienced the following emotions while playing the game.

115. Interested *

Markeer slechts één ovaal.

	Not at all
1	◯
2	◯
3	◯
4	◯
5	◯
	Extremely

116. Excited *

Markeer slechts één ovaal.

	Not at all
1	◯
2	◯
3	◯
4	◯
5	◯
	Extremely

117. Strong *

Markeer slechts één ovaal.

	Not at all
1	◯
2	◯
3	◯
4	◯
5	◯
	Extremely

118. Enthusiastic *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

119. Proud *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

120. Alert *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

121. Inspired *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

122. Determined *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
Extremely

123. Attentive *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
Extremely

124. Active *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
Extremely

Negative Affect subscale (PANAS-N)
Rate how much you have experienced the following emotions while playing the game.

125. Distressed *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

126. Upset *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

127. Guilty *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

128. Scared *

Markeer slechts één ovaal.

Not at all
1
2
3
4
5
Extremely

129. Hostile *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
Extremely

130. Irritable *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
Extremely

131. Ashamed *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
Extremely

132. Nervous *

Markeer slechts één ovaal.

Not at all
1 ◯
2 ◯
3 ◯
4 ◯
5 ◯
Extremely

133. Jittery *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Extremely

134. Afraid *

Markeer slechts één ovaal.

Not at all

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Extremely

The NASA Task Load Index (NASA-TLX)
This part of the questionnaire measures the cognitive load of the game.

135. How mentally demanding was the game? *

Markeer slechts één ovaal.

Very low

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

6 ◯

7 ◯

8 ◯

9 ◯

10 ◯

Very high

136. How physically demanding was the game? *

Markeer slechts één ovaal.

Very low
1
2
3
4
5
6
7
8
9
10
Very high

137. How hurried or rushed was the pace of the game? *

Markeer slechts één ovaal.

Very low
1
2
3
4
5
6
7
8
9
10
Very high

138. How successful do you think you were in accomplishing the goals of the game? *

Markeer slechts één ovaal.

Very low

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

6 ⬭

7 ⬭

8 ⬭

9 ⬭

10 ⬭

Very high

139. How hard did you have to work to accomplish your level of performance? *

Markeer slechts één ovaal.

Very low

1 ⬭

2 ⬭

3 ⬭

4 ⬭

5 ⬭

6 ⬭

7 ⬭

8 ⬭

9 ⬭

10 ⬭

Very high

140. How insecure, discouraged, irritated, stressed, and annoyed were you during the game? *

Markeer slechts één ovaal.

	Very low
1	◯
2	◯
3	◯
4	◯
5	◯
6	◯
7	◯
8	◯
9	◯
10	◯
	Very high

Additional notes

141. What are your thoughts on using facial expressions as a game mechanic in the game that you've played? *

L Follow-up questionnaire

Follow-up questionnaire

You have filled in a questionnaire about two versions of a game you played. Some results observed in that questionnaire raised some questions. This follow-up questionnaire is to find an explanation for the observed results. please keep the answers around 2-3 sentences.

* Verplichte vraag

1. Could you name some advantages of the version with facial expressions? *

2. Could you name some disadvantages of the version with facial expressions? *

3. Could you name some advantages of the version without facial expressions? *

4. Could you name some disadvantages of the version without facial expressions? *

5. What do you think made the version with facial expressions less usable? *

6. Do you feel like the version with facial expressions required more effort, if so, why? *

7. Do you feel like the version with facial expressions was more physically demanding, if so, why? *

8. Do you feel like the version with facial expressions was more challenging, if so, why? *

9. Do you feel like the version with facial expressions made you feel more negative feelings, if so, why? *

10. Could you think of an adjustment to the version of the game with facial expressions that would make it more fun and engaging to use than the traditional version?