

Applying Facial Expression Analysis Software to Estimate the User Experience of LEGO building by Analyzing Facial Action Units

Lisa Bondo Andersen
Aalborg University
Department of Electronic Systems
Aalborg, Denmark
lan16@student.aau.dk

Mathias Thygesen
Aalborg University
Department of Electronic Systems
Aalborg, Denmark
mthyg19@student.aau.dk

Abstract

This study explores the application of Facial Expression Analysis (FEA) to estimate the user experience (UX) during LEGO building activities by analyzing Facial Action Units (AU). Traditional user experience (UX) research often relies on subjective self-reports, which are subject to biases. In contrast, we attempt to employ FEA to provide a more objective and quantifiable measure by capturing facial expressions. We utilized OpenFace 2.0, a tool for detecting facial landmarks and recognizing AUs, to analyze participants' facial expressions as they engaged in a LEGO building task. The study involved a remote experiment setup with 18 participants who built a LEGO set at home while being recorded on video. These videos were processed to extract AU data, which were then compared to participants' self-reported ratings of enjoyment, frustration, challenge, boredom, and excitement on a 7-point scale.

Despite extensive data collection and analysis, the machine learning models trained, based on Random Forest regression and classification, to predict subjective assessments from AU data, showed poor performance. This may be due to the solitary nature of the task affecting participants' facial expressiveness or insufficient variation in self-reported experiences. The findings suggest that while FEA holds potential for UX research, further methodological refinements are needed. Future research should consider more frequent self-assessments, deploying more complex machine learning models, and potentially incorporating additional biometric measures to improve accuracy of emotional state predictions.

Keywords: User Experience, Subjective Assessments, Emotions, Facial Expression Analysis, Machine Learning, Facial Action Units, OpenFace

ACM Reference Format:

Lisa Bondo Andersen and Mathias Thygesen. 2024. Applying Facial Expression Analysis Software to Estimate the User Experience of LEGO building by Analyzing Facial Action Units. In *Proceedings of Engineering Psychology (Master's Thesis)*. ACM, New York, NY, USA, 14 pages.

Master's Thesis, May 31 2024, Aalborg University
2024.

1 Introduction

When measuring UX, the goal is to understand and interpret user perceptions. Traditional methods like questionnaires and interviews are, however, prone to biases (Orne, 2002; Wetzel et al., 2016; Longo, 2018). To gain more objective insights, recent research within the field is attempting to implement biometric measures to gain a more objective and reliable image of participants' cognitive state in terms of e.g. work load, attention and emotions (Santoso et al., 2016; Čertický et al., 2019; Bañuelos-Lozoya et al., 2021). However, sensors measuring e.g. electrocardiogram (EEG) or galvanic skin response (GSR) often disturb the user's interaction with the product because they have to be attached to the user to be able to measure data. This limits the user's movements of their head or hands. FEA, on the other hand, only requires recordings of participants' face to output useful data, which makes this method one of the least intrusive biometric measuring methods, thus maintaining a higher level ecological validity (Orne, 1962).

FEA is often seen used to measure emotions, because facial expressions can provide valuable information about users' social and affective states, through a universal form of non-verbal communication (Baltrušaitis et al., 2016; Chang et al., 2024; Paul Ekman Group, 2024). Reflections of a person's emotional state can be present on the face even if we try to conceal it, in the form of microexpressions occurring in half a second or less, which however, can make this type of communication difficult to detect (Paul Ekman Group, 2024).

Prior research regarding the connection between self-assessed emotions and facial expressions show correlations between positive and negative emotions in connection with purchase intent in sensory tests (Samant and Seo, 2020). A correlation has also been found between the two for arousal and valence in a nonverbal driver-pedestrian interaction (Rao et al., 2023), as well as regarding overall subjective experiences of players in an online video game tournament (Mavromoustakos-Blom et al., 2021).

1.1 Facial Action Units

The facial action coding system (FACS) was developed by Ekman and Friesen (1978), who formalized AUs into a system of

44 AUs. The system was created to include all distinguishable facial movements into distinct categories, not only related to emotion-specific movements, but rather focused on any anatomically possible facial movements (Ekman and Friesen, 1978). The system was created in a way where multiple muscles are included in a single AU if these are activated simultaneously. This means that there is no direct correspondence between single muscles and AUs, since the same muscles can also be seen included in different AUs which is the case for *Inner Brow Raiser* (AU1) and *Outer Brow Raiser* (AU2).

This study implements OpenFace 2.0 for FEA, an open source toolkit for conducting FEA through the automatic detection of activation in AUs (Baltrušaitis et al., 2018; Chang et al., 2024). The tool offers functionalities such as facial landmark detection, head pose estimation, eye gaze tracking, and AU recognition (Baltrušaitis et al., 2018). When receiving an image as an input, the tool will detect whether or not a face is present, followed by using facial landmark detection to identify key points on the face, such as the corners of the eyes and the edges of the lips. These landmarks are e.g. used to estimate head pose and eye gaze. The tool is able to detect both the presence and intensity of various AUs (Baltrušaitis et al., 2016; Baltrušaitis et al., 2018) with Table 1 showing which AUs OpenFace 2.0 detects.



















1.2 Literature Review

To further investigate the use of FEA in research and to gain inspiration for how the experiment of this study could be structured, a narrative review was carried out, as per recommendations by Snyder (2019) and Ferrari (2015). For the review, we decided to use the two databases *APA PsycInfo* and *Scopus* as well as to only include articles and conference papers, meaning that e.g. book chapters and user manuals on how to conduct experiments using FEA were not included, because we wanted to gain inspiration from methods used in prior research.

The review consisted of three rounds as outlined in Table 2. The first round aimed to gain an overall understanding of how and when FEA was used in research. In the second round, we employed more well-defined search prompts and inclusion criteria based on insights from the first round. Specifically, papers on medical research were excluded in the second round, as they typically focused on measuring pain through facial expressions or comparing expressions between neurotypical and neurodivergent individuals, which were deemed unrelated to the scope of this study. In the first two rounds, abstracts were reviewed to decide whether to include or reject the papers, and in the third round, the entire articles of the papers included from the second round were reviewed. 3 papers were excluded in this round due to them being inaccessible or written in foreign languages.

Contexts of FEA in Prior Research. The review showed that FEA was used in various different contexts involving

Table 1. The AUs that OpenFace 2.0 detects (Baltrušaitis et al., 2018). Predictions are available for intensity and presence for all AUs, except for AU28 where only presence is predicted.

AU	Full name	Illustration
AU1	Inner brow raiser	
AU2	Outer brow raiser	
AU4	Brow lowerer	
AU5	Upper lid raiser	
AU6	Cheek raiser	
AU7	Lid tightener	
AU9	Nose wrinkler	
AU10	Upper lip raiser	
AU12	Lip corner puller	
AU14	Dimpler	
AU15	Lip corner depressor	
AU17	Chin raiser	
AU20	Lip stretched	
AU23	Lip tightener	
AU25	Lips part	
AU26	Jaw drop	
AU28	Lip suck	
AU45	Blink	

different kinds of products/stimuli, with one of the most common contexts being sensory tests using odor (Savela-Huovinen et al., 2021), food samples (Gülşen et al., 2023) and drink samples (Matsufuji et al., 2023; Crist et al., 2018; de Wijk et al., 2021; Zhi et al., 2020; Samant and Seo, 2020; Samant et al., 2017; Zhi et al., 2018) as stimuli. The second most common context was using videos, e.g. ads (Walsh et al., 2017; Holiday et al., 2023; Zeng and Lobo Marques, 2023), public service announcements (PSA; Parvanta et al., 2022; Hammond et al., 2022), emotional videos (Kassas et al., 2022; Zarei et al., 2022), video lectures (Rodríguez-Fuertes et al., 2022), and Instagram posts (Sass and Fekete, 2022) as stimuli. One of these was Parvanta et al. (2022), who distributed an online survey with three 30-second anti-smoking videos and

Table 2. The search prompt, inclusion criteria and rejection criteria as well as the number of abstracts/papers read and included in each round of the literature review. In the second round, 3 versions of the search prompts were used. "Facial expression analysis" was replaced with "facial expression recognition" and "facial action units" to see if these variations produced different results in the databases.

Round	Search Prompt	Inclusion Criteria	Rejection criteria	Read	Included
1st	"Facial expression analysis"	1. Measures facial movement in relation to a psychological response	1. Review articles 2. Development of FEA	99	59
2nd	1. "Facial expression analysis" AND ("scale" OR "survey" OR "interview") NOT "pain" 2. Alternatively: a) "facial expression recognition" b) "facial action units"	1. Exposure to stimuli 2. Measures facial movement related to mood, emotion etc. compared with self-reported subjective assessment	1. Medical research unrelated to emotions	202	33
3rd	-	1. Included papers from the 2nd round	1. Not written in English/Danish 2. No access	33	30

multiple choice questions with a platform set up to record the participants' faces while watching the videos. Through the survey, the participants assessed the effectiveness of the PSAs, their own desire to stop smoking and to share the videos on social media platforms, as well as their willingness to complete a *Quit Now Trial*.

In two of the reviewed articles, FEA was used to assess the experience of tangible products, specifically packaging material (Clark et al., 2021; López-Mas et al., 2022). Clark et al. (2021) measured product-associated emotions using a check-all-that-apply sheet and an association test, where the participants e.g. were presented with images of two milk packaging samples with a neutral or emotional word between them. The participants were then asked to choose between the packaging samples and to categorize them using the presented word. During the test, the participants also simulated themselves inspecting the packaging in a grocery store. Product-associated emotions, product acceptability and purchase intent were measured after the interaction and later compared with the participants' facial expressions during the test.

Other less common contexts involving FEA all included online interactions, such as chatbot interactions (Carmichael et al., 2021), video game tournaments (Mavromoustakos-Blom et al., 2021; Jones et al., 2021), and online shopping experiences (Mookherjee et al., 2021). These contexts are comparable to the ones using videos as stimuli, since in all cases, stimuli were presented on a PC. FEA was seen less used in contexts involving physical experiences using a football game (Richlan et al., 2023), a driver/pedestrian interaction (Rao et al., 2023), or an opera performance (Ceccacci et al., 2023) as stimuli. Lastly, we found that FEA was used to measure facial expressions during self critique (Halamová et al., 2023), looking at different chart types indicating energy consumption (Kremsner et al., 2023), and in a corporate social responsibility scenario (Deng et al., 2023).

Self-assessments and FEA. When evaluating the effect of FEA, results were often compared to self-assessment parameters. While these, as priorly described, are subject to biases, various methods have been found applied in literature to carefully collect self-reports and reliably compare these with FEA data. The self-reported parameters were usually emotions, also including arousal and valence (e.g. Richlan et al., 2023; Deng et al., 2023; Clark et al., 2021; Walsh et al., 2017), and less commonly seen was engagement (Holiday et al., 2023), desire to stop smoking (Parvanta et al., 2022), attention (Hammond et al., 2022), self-criticism (Mookherjee et al., 2021), regret and disappointment (Halamová et al., 2023), purchase intent (López-Mas et al., 2022), and liking (Zhi et al., 2018; Zhi et al., 2020). A combination of emotions and engagement were measured in one study (Rodriguez-Fuertes2022), while a combination of emotions and liking were measured by de Wijk et al. (2021) and Samant et al. (2017). Data regarding these parameters were most often collected through a scale either presented in a questionnaire or an interview during or after the experiment.

1.3 Structure of the Study

In the literature review, only two papers were found using tangible products as stimuli, where Clark et al. (2021) used milk packaging samples and López-Mas et al. (2022) used packaging for fish burgers, meaning that no papers were found focusing on the entire product experience, including both the packaging and the interaction with the product itself. Thus, FEA is rarely seen used in a context of evaluating the UX of an interaction, much less in the case of a tangible product, and never with these combined in a remote experiment. Based on the inspiration gained from the literature review as well as the limitations of traditional UX methods, this study aims to expand research within FEA to include methods for measuring UX of a tangible product in a remote

experimental setup. This addresses a gap in the literature where no research currently targets this specific topic.

Our study aims to collect and analyze data on participants' facial expressions in a LEGO building context in order to use the data to investigate the relation between facial movements (AUs) and self-reported assessments of *enjoyment*, *frustration*, *challenge*, *boredom*, and *excitement*. Pending a correlation identified between these, the aim is furthermore to investigate how machine learning models can be trained based on AUs to predict the subjective assessments. This study was conducted in collaboration with The LEGO Group, who graciously provided the LEGO sets used as stimuli in the experiment.

2 Measuring Emotions with FACS

Research shows general agreement of which AUs are active for the 6 basic emotions; happiness, sadness, anger, fear, surprise and disgust (Ekman and Friesen, 1978; Tejada et al., 2022; Sharma et al., 2022). This, however, is not the case for the 5 parameters of this study, which are less researched in literature. Table 3 shows which AUs are expected to be active when someone experiences high levels of each parameter, and which emotions have been shown in prior research to result in similar expressions as the basic emotions.

Table 3. The 5 parameters used in the experiment, and which AUs these activate as well as which AUs are activated by similar emotions. *Happiness** indicates that the AUs active for this emotion are based solely on assumptions due to the lack of research in the field.

Parameter	Similar emotion	AU active	Reference
Enjoyment		AU6, AU12	Gosselin et al., 2002
	Happiness	AU6, AU12	Ekman and Friesen, 1978
Frustration		AU12, AU43	McDaniel et al., 2007
	Anger	AU4, AU5, AU7, AU23	Ekman and Friesen, 1978
	Sadness	AU1, AU4, AU15	Ekman and Friesen, 1978
Challenge		-	
Boredom		AU4, AU7, AU12	Sharma et al., 2022
	Neutral	none	McDaniel et al., 2007
Excitement		-	
	Happiness*	AU6, AU12	Ekman and Friesen, 1978

Enjoyment is seen to have a correlation with experiencing happiness which should result in an activation of AU6 and AU12 (Gosselin et al., 2002). For *excitement*, no literature was found to show correlation with any of the basic emotions, but we made the assumption that happiness would also correlate with this emotion. *Frustration* is shown to activate AU12 and AU43 (Sharma et al., 2022; McDaniel et al., 2007), but OpenFace does not measure AU43 (eye closure), so to not confuse *frustration* with *enjoyment* and *excitement*, the AUs activated by anger and sadness were considered, as *frustration* has been reported to include these feelings (Huntington,

2024; Abler et al., 2005). *Challenge* was not found to activate any specific AUs nor to have a correlation with any of the basic emotions. *Boredom* was found by Sharma et al. (2022) to activate AU4, AU7 and AU12 and by McDaniel et al. (2007) to activate no AUs, similar to that of a neutral facial expression. In this regard, Sass and Fekete (2022) also describe that facial expressions of *boredom* are easily influenced by an observational situation, and might be concealed in such a situation.

3 Method

An experiment was conducted to collect data on participants' facial expressions and subjective assessment of the 5 parameters; *enjoyment*, *frustration*, *challenge*, *boredom*, and *excitement*. The experiment consisted of a remote experiment setup where the participants had to build the LEGO Set 10313 called *Wildflower Bouquet* while recording themselves. All material required to complete the experiment, being an instruction document, the LEGO set, and a QS, were delivered to the participants in order for them to complete it at their own pace when they had time within a specified period of 4 weeks. We chose a remote setup for the experiment to increase the ecological validity, since people who buy a LEGO set will typically build it at home. Participants were not told that the purpose of them being recorded was to analyze their facial expressions because this could affect their behavior including their facial expressions during the experiment (Sass and Fekete, 2022).

3.1 Participants

All participants were recruited through a recruitment questionnaire. 20 people (9 women and 11 men), all in the age group 18-30 years old, participated in the experiment, while applicable data was received from 18 participants. None of the participants needed to wear glasses during the experiment or had a excessive facial hair, which were found by Clark et al. (2021) and Crist et al. (2018) to cause problems when using FEA software. Furthermore, none of the participants had built set 10313 before, which was important since people who had built the set before might not be biased towards the 5 parameters as it would not be a new experience for them. The participant group consisted of both people who were experienced LEGO builders and people who had never built with LEGO before.

3.2 LEGO Set

We chose to include LEGO set 10313 (see Figure 1) as the stimuli in the experiment. This specific set was chosen as it was, during the time of this study, one of the most popular LEGO sets among LEGO builders as well as non LEGO builders. Furthermore, the set is targeted at an adult age group of 18+, meaning that we were able to recruit participants in this age group instead of focusing on children, who are the typical

target group of LEGO, since it would be difficult to carry out this type of experiment using children as participants. The set includes 939 pieces, distributed in enumerated bags from 1-4. In the set, there are 8 species of flowers with some of them appearing up to 3 times, for a total of 17 flowers in the set.

Figure 1. The LEGO Set 10313 called *Wildflower Bouquet* that the participants built during the experiment.



3.3 Experiment Material

An instruction document was made and handed to the participants explaining their task during the experiment, being that they had to build 17 flowers in total and assess 5 scales (one for each of the chosen parameters) after building each flower, and that they should set aside approximately 2 hours to complete it. The document also included a link to an instruction video explaining the materials they needed as well as how to set up their recording device to ensure optimized data collection. The video explained the entire experiment procedure, including the unboxing, the building process, and the debriefing. A pilot study was carried out prior to conducting the experiment, where the introduction video was refined.

The QS included a checklist that participants had to answer yes to before starting the experiment with the following statements:

- Are you alone in the room?
- Is it quiet in the room (no music, other people talking etc.)?
- Have you removed your glasses (if you were wearing any)?

- Have you removed your hair from your face by putting it behind your ears or putting it up in a ponytail (if you have long hair)?
- Do you have at least 10 GB of space free on the device that you will use to record?
- Did you put your phone on 'do not disturb' or 'flight mode' (if you're using your phone to record)?
- Do you have at least 80% battery power on your phone (if you're using your phone to record)?

On the same page, five 7-point scales for each of the 5 parameters were presented for the participants to assess their general mood before the experiment. Furthermore, the QS included a page with an image and the same five 7-point scales for the unboxing, for each of the 17 flowers, and for the overall experience. Here, the images were of the LEGO box, of each flower as shown in the LEGO manual, and of all the flowers assembled into a bouquet, respectively. All pages with scales included a description of what they were assessing, where e.g. pages related to the flowers included the following instruction: *"Rate the following based on the steps you followed to build [Flower]"*.

The order of the 5 parameters, the participants had to assess, were balanced using a 5x5 latin square to reduce carry-over effects, resulting in 5 different versions. The order of the parameters remained the same for each participant throughout the QS, as it was assumed to be easier for the participants, since they had to assess each parameter 20 times during the experiment. We considered counterbalancing the order of flowers as well, but refrained from doing so, as it would presumably decrease the ecological validity to change the original manual. Further, since each of enumerated bags include pieces for 2-7 flowers each, it would only allow for a very limited set of versions.

The last page of the QS presented the debriefing, where the participants were asked about their experience with participating and building the LEGO set, including the moments they found most enjoyable, most frustrating, most challenging, most boring, and most exciting. These questions were to be answered out loud as the final part of the recording. This was the only part of the experiment where the participants were allowed to talk, as talking can interfere with FEA (Shah et al., 2013).

Experiment Procedure. When the participants had received the experiment material, they first read the instruction document and watched the instruction video to prepare for the recording and the experiment procedure. The participants completed the experiment by following the instructions of the video, and ensuring they could affirm the statements on the first page of the QS, after which they reported their general mood before opening the box and assessing the 5 parameters based on this. Next, the participants followed the LEGO instruction manual chronologically, by building one flower at a time and assessing the 5 parameters based on

the flower they had just built. After having built all 17 flowers, the participants assessed the 5 parameters based on their overall experience of participating in the experiment and answered the questions in the debriefing in either English or Danish. After completing the experiment, the participants sent the recording and the assessments from the QS back to us.

4 Results and Analysis

Over the span of 4 weeks, we collected videos and parameter ratings from 18 participants. As each participant submitted their recording and QS, we continuously processed and analyzed the videos using the OpenFace software. This section evaluates the suitability of the collected data for model training, focusing on whether the outputted AUs can predict users' self-assessments of the 5 parameters.

4.1 Data Preparation

Prior to running the videos through the OpenFace software, videos were meticulously edited to solely include segments of the relevant 18 steps, being the unboxing and building of the 17 flowers. Excluded segments include the times when participants were filling out the QS, taking breaks, and unpacking plastic bags. The trimmed videos were then sped up by 300%, ensuring a frame rate of ≥ 10 frames per second to still capture micro-expressions (Paul Ekman Group, 2024). This acceleration allowed for efficient processing while maintaining the integrity of the expressions recorded. The editing process significantly reduced the total video length from 39 hours and 42 minutes to a more manageable 13 hours and 14 minutes.

Participants provided self-assessments for the parameters; *enjoyment*, *frustration*, *challenge*, *boredom*, and *excitement*. As seen in Figure 2, *Enjoyment* and *excitement* consistently received the highest ratings, with mean values typically ranging between 4 and 6 on the 7-point scale. In contrast, *frustration*, *challenge*, and *boredom* were rated lower, with mean values generally between 1 and 4. These ratings were consistent across all steps, with periodic spikes in challenge and frustration during more complex building steps.

Participants spent varying amounts of time on building the LEGO set, ranging between 90 and 180 minutes, as well as on the different steps of the building process. The least amount of time was spent on unboxing, while the most time-consuming steps were some of the last ones, specifically F13, F14, and F15.

4.2 OpenFace Analysis

The OpenFace toolkit was employed to analyze participants' facial expressions. The FeatureExtraction command was utilized to generate *Histogram of Oriented Gradients* (HOG) files and videos that identified facial features. Additionally, CSV files were produced, containing x and y coordinates of

landmarks as well as the presence and intensity values of 18 AUs for each frame analyzed.

For the CSV output files, post-processing was performed prior to feeding the data to the models. Frames with *confidence* levels of facial detection below 0.85 were excluded, as well as frames with no face detected. This resulted in a 2.20% reduction of the total dataset, leaving 1,035,918 data points for further analysis. The remaining data points underwent normalization for intensity values to account for individual differences in expressiveness. In the output from the OpenFace analysis, intensity values ranged from 0-5, but were normalized to range from 0-1 to match the range of the presence variables. The distribution of AU intensity across all steps is presented in Figure 3.

4.3 Model training

Table 4 provides a comprehensive overview of the different methods used to format data for training machine learning models. These methods involve processing both the intensity (I_{AU}) and presence (P_{AU}) values of AUs as detected by the OpenFace software. This was done as part of the model fitting, to find the combination of approaches for best model performance.

Regression and Classification Models. Five models were trained using random forest regression and classification techniques in Python (V3.12.2) with the following commands from the *scikit-learn* package (Scikit-learn, 2024a; Scikit-learn, 2024b):

```
RandomForestRegressor(random_state=42)
RandomForestClassifier(random_state=42)
```

Random forest is a commonly used method for initial model training due to its relatively high interpretability, particularly in comparison to methods such as neural networks, as it provides feature importance metrics (Breiman, 2001). Additionally, Random Forest is capable of handling complex interactions between variables, unlike linear models (Cutler et al., 2007), and due to its approach of combining multiple decision trees, it is resistant to overfitting (Biau, 2012). However, this method is computationally intensive, particularly when applied to large datasets (Probst et al., 2019; Bergstra and Bengio, 2012). Moreover, Random Forest classification supports multi-category classification, whereas some alternative methods are restricted to binary classification (Hastie et al., 2009). This restriction could result in a loss of data granularity, as categories would have to be simplified into binary outcomes, such as positive and negative, for each self-assessment parameter.

The features for the models were based on the intensity and presence of AUs, and the target variables were the 5 parameters. The models were trained using several combinations of the AU intensity and presence approaches indicated in Table 4. Table 5 shows performance metrics of regression

Figure 2. Mean ratings for each parameter for each step. Error bars show 95% CI.

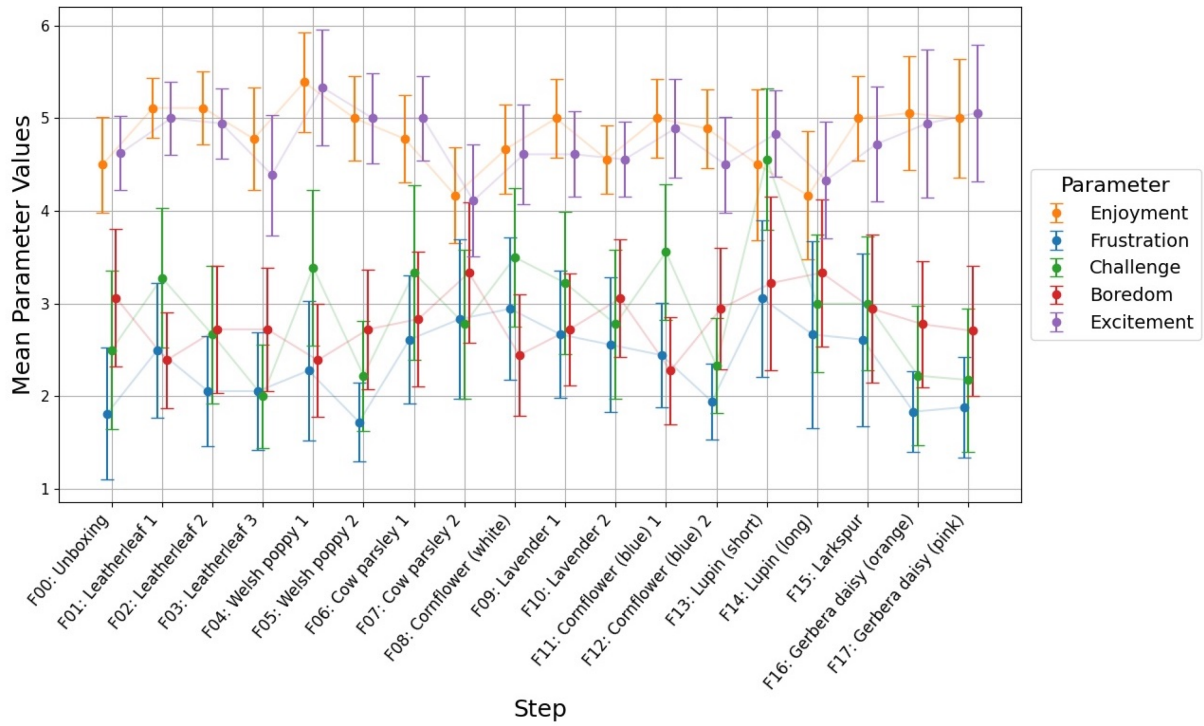


Figure 3. Mean *intensity* for the 17 AUs for each step, normalized to range from 0-1.

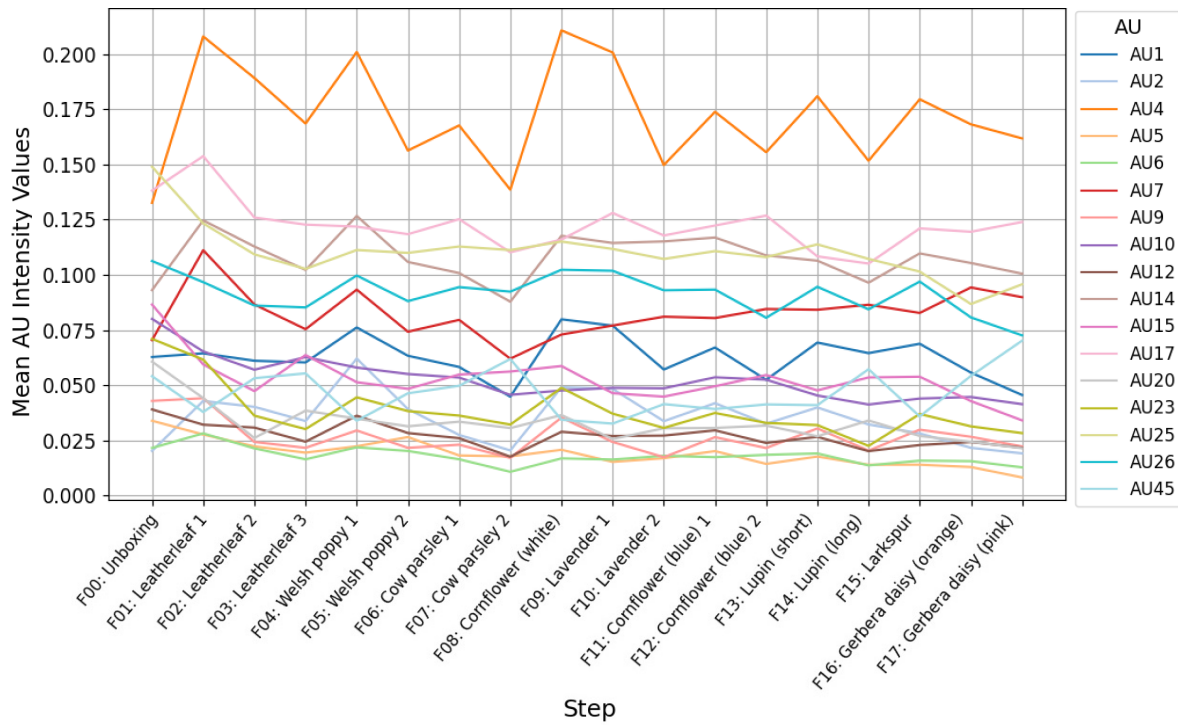


Table 4. The different approaches to formatting data that were used to train the machine learning models. The formatting was done by processing AU *intensity* (I_{AU}) and *presence* (P_{AU}) values from the OpenFace output.

AU type	Term	Range	Mathematical Explanation	Description
Intensity	Normalized by max AU	0-1	$I_{normMax} = \frac{I_{AU}}{5}$	Intensity of AUs normalized by dividing with 5, the max value any AU can reach.
Intensity	Normalized by individual participant	0-1	$I_{normInd} = \frac{I_{AU}}{I_{indMax}}$	Intensity of AUs normalized by dividing with the max value of the given AU for the given participant.
Intensity	Mean overall	0-1	$I_{meanOverall} = \frac{1}{N} \sum_{i=1}^N I_{AU_i}$	Intensity mean value per step per participant for “normalized by max AU” values.
Intensity	Mean if present	0-1	$I_{meanPresent} = \frac{1}{N_p} \sum_{i=1}^{N_p} I_{AU_i}$	Intensity mean value per step per participant for “normalized by max AU” values excluding 0 values, so only including intensity values when the AU is present.
Presence	Mean	0-1	$P_{mean} = \frac{1}{N} \sum_{i=1}^N P_{AU_i}$	How often AU is present per step per participant.
Presence	Significant from binomial distribution	0, 0.5, 1	$P_{sig} = \begin{cases} 1 & 0.5 < CI_{95\%} \\ 0.5 & 0.5 \in CI_{95\%} \\ 0 & 0.5 > CI_{95\%} \end{cases}$	If presence mean is significantly more or less present than not per step per participant with 95% CI.
Presence	Dynamics	0-1	$P_{dynamics} = \frac{\text{Number of changes}}{N-1}$	How often AU changes between being present and not present.

and classification models trained for all 5 parameters, using $I_{meanOverall}$ based on $I_{normInd}$ data and P_{sig} data. Because of dependency in our data, as multiple data points come from the same participant, we decided to exclude a set of random participants from the training set instead of a typical random split. Initially, we assigned data from P6, P15 and P18 (randomly selected) to the test set, but also attempted other variations, resulting in no substantial improvements in model performance.

A common rule in machine learning is to have 10 times, or preferably 100 times, as many data points as the number of features (Maxwell et al., 2018). With only 227 data points in the training set, this would allow for fewer features than we had available, and even with attempting 10, 15, and 20 of the most important features, the model performance still did not improve. Due to the poor performance, after attempting several combinations of intensity and presence of AUs, a different approach was tried out. This approach involved using the entire data set of individual values, and assigning the parameter values to each data point in a given step for a given participant, resulting in approximately 900,000 data points in the training set. The performance metrics for the models trained on these data, using intensity data normalized by individual participant, can be seen in Table 6. For this, RandomizedSearchCV from the *scikit-learn* package was applied to reduce the computing time. Once again, the metrics show quite poor performance for both regression and classification, also after attempting several variations of excluded participants, hyperparameters etc.

Table 5. Evaluation metrics for the 10 random forest models, 2 for each of the 5 self-assessment parameters. MAE: Mean absolute error; MSE: mean squared error; RMSE: root mean squared error; Acc: accuracy; prec: precision; rec: recall.

Models trained on mean data points								
Target	Regression				Classification			
	MAE	MSE	RMSE	R ²	Acc.	Prec.	Rec.	F1
Enjoyment	1.05	1.88	1.37	-0.14	0.31	0.14	0.16	0.15
Frustration	1.3	2.33	1.53	-0.09	0.31	0.08	0.13	0.10
Challenge	1.42	2.68	1.64	0.06	0.24	0.17	0.19	0.15
Boredom	1.57	3.23	1.8	-2.78	0.13	0.15	0.11	0.09
Excitement	0.88	1.29	1.13	-0.43	0.31	0.20	0.23	0.21

Table 6. Evaluation metrics for the 10 random forest models, 2 for each of the 5 self-assessment parameters. MAE: Mean absolute error; MSE: mean squared error; RMSE: root mean squared error; Acc: accuracy; prec: precision; rec: recall.

Models trained on individual data points								
Target	Regression				Classification			
	MAE	MSE	RMSE	R ²	Acc.	Prec.	Rec.	F1
Enjoyment	1.21	2.62	1.62	-0.12	0.36	0.12	0.14	0.11
Frustration	1.43	2.97	1.72	-0.01	0.27	0.13	0.12	0.09
Challenge	1.56	3.10	1.76	-0.12	0.19	0.17	0.17	0.15
Boredom	1.57	3.22	1.79	-2.89	0.17	0.11	0.11	0.09
Excitement	0.87	1.22	1.10	-0.50	0.44	0.15	0.19	0.15

4.4 AU Intensity Compared with Self-reports

The visual representations in Figures 4 and 5 highlight the discrepancies between actual self-assessment values and the values assumed from literature. AUs indicated with bold lines were expected to increase with higher self-assessment ratings. Note that the data set is quite skewed, meaning there are few data points for the low values of *enjoyment* and *excitement* and vice versa for *frustration* and *boredom*.

5 Discussion

The study reveals challenges in using facial expression data to predict subjective user experiences in a solitary LEGO building task. The findings suggest that while a remote test setup is certainly suitable for recording and analyzing facial expressions during LEGO building, identifying a correlation

between activation of certain AUs and the 5 self-assessment parameters has proven challenging.

Despite extensive parameter tuning, the models demonstrated limited success in accurately predicting the self-assessment parameters. The performance metrics indicated low MSE and accuracy scores, suggesting that the collected facial expression data did not sufficiently capture the nuanced emotional experiences of the participants. As indicated by Figure 4 and 5, the AUs did not activate as expected with regards to the self-assessed parameters. We suggest three explanations for this. Firstly, the relationship between the five parameters and the AUs described in the literature might not hold true in the specific context of this study. Prior research has primarily been on the 6 basic emotions, meaning that some assumptions had to be made, which may have been faulty. Secondly, the solitary environment could have caused the participants' facial expressions to remain neutral

Figure 4. Activation of AUs depending on the participants' *enjoyment* and *excitement* ratings. The bold lines indicate the AUs that were assumed to increase with increased ratings.

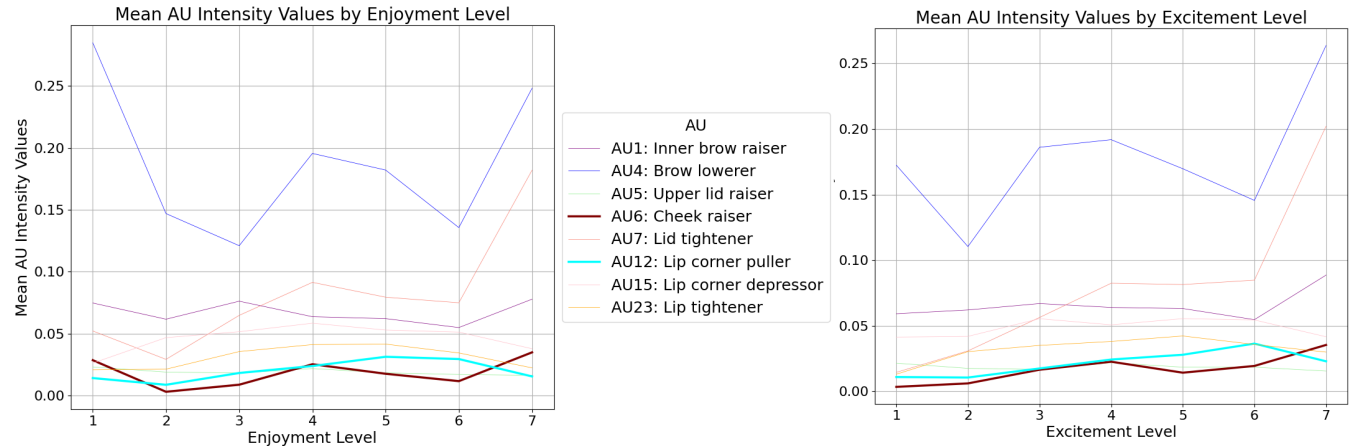
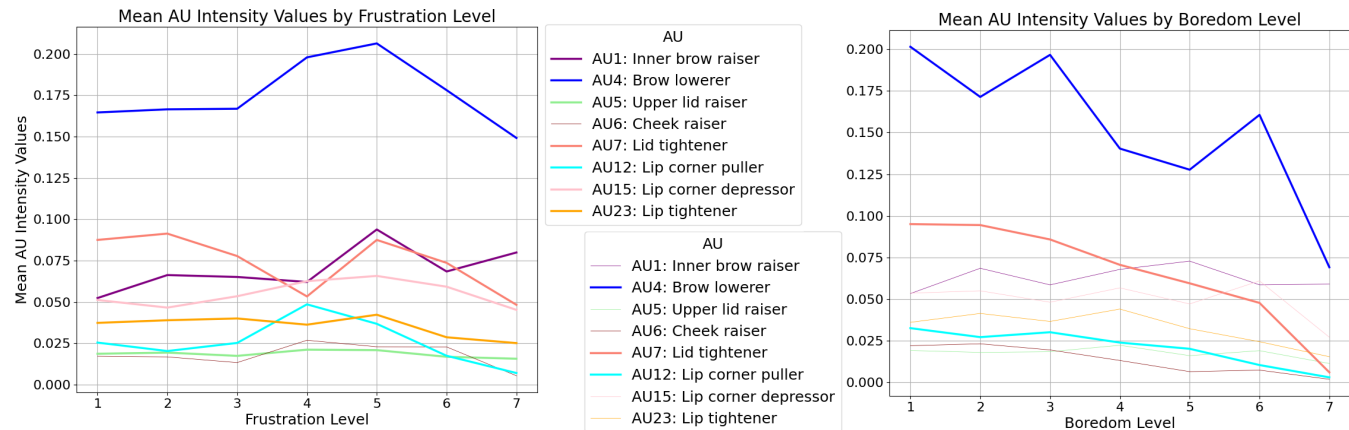


Figure 5. Activation of AUs depending on the participants' *frustration* and *boredom* ratings. The bold lines indicate the AUs that were assumed to increase with increased ratings.



even though they were experiencing e.g. *enjoyment*. This is supported by Fridlund (1991), who found that some aspects of emotional expressions, and especially those of joy, are highly socially dependent. Thus, while the remote setup resulted in higher ecological validity (Orne, 1962), it might have negatively affected the intensity and presence of facial expressions. Future work could attempt to research experiment setups with building LEGO as a social activity, although this introduces a variety of new issues, for instance whether the participants' facial expressions are affected by the stimuli presented in the experiment or are due to a social connection between the participants. Thirdly, while the expressions of the 6 basic emotions have shown to be consistent across cultures, genders, and age groups (Ekman and Friesen, 1978; Tejada et al., 2022), the 5 parameters of this study might be subject to individual differences. This variability was evident in the videos, where some participants expressed what appeared to be concentration by sticking their tongue out, while others lowered their brows.

Additionally, as presented in Figure 2, participants generally did not differ much in their ratings within each parameter across the 18 steps. This implies that the experiment did not sufficiently affect participants' experiences, leading us to have little data for e.g. high values of *frustration*. The skewed data distribution is problematic for model training, as it results in an insufficient number of data points at the extreme ends, leading to poor model performance. Even though, we attempted to present the participants with a rather complex LEGO set, it did not result in much variation. This is also evident in the mean intensity of AUs, as seen in Figure 3, where no steps seem to consistently activate some AUs more than others. Evaluating experiences with shorter time intervals is likely to reduce both of these issues and lead to more variation in data. This way, for instance 20 seconds with high levels of frustration would be apparent in data, unlike the current setup where it gets filtered out in an assessment based over possibly 10 minutes, where the remaining time might be neutral or enjoyable.

Shifting from aggregated mean values to individual data points showed only marginal improvements in model performance. This indicates that directly assigning the ratings for the 5 parameters is insufficient; instead, each data point must robustly reflect the participant's emotional state at the given time. Thus, the current approach might have inaccurately assigned high self-assessments to largely neutral facial expressions, leading to compromised data quality. An approach to enriching the data with more ratings would be to ask the participants to rate the parameters with higher frequency. Practically, we do not consider it possible to probe them with high enough frequency during a remote study, especially not considering the duration of this experiment. Instead, conducting shorter experiments with a larger sample size, for instance by having participants build one flower each, would

offer more room to probe them with higher frequency. Moreover, the approach by Čertický et al. (2019) could be adopted. In a similar setup, although with researchers present for the experiment, they retrospectively watched the recordings with the participants and asked them to rate their level of enjoyment on a 5-point scale during different time intervals. Reaching a larger sample size would likewise enhance the model, as the current dataset, although substantial in size, is vulnerable to overfitting due to the low number of participants and thus limited variability in facial expressions.

6 Future Work and Final Remarks

With the current study being the only one of its kind utilizing FEA in UX research in the context of interacting with a tangible product in a remote experiment setup, we contribute with valuable insights, useful for future work within the field.

Implementing the methodological enhancements of more frequent self-assessments at smaller intervals, as proposed in the previous section, would provide a stronger foundation for training an effective model. Additionally, future work could involve refining the model by exploring different approaches such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), which are suitable for image/video analysis and processing of time-series data (Bishop, 2006, p. 267-269; Petneházi, 2019). The current models do not account for the sequence in which certain AUs are presented, which CNNs and RNNs can handle effectively. Although an attempt was made to implement dynamics in the presence of AUs, this proved insufficient. However, the dynamics data were among the most important features in all trained models, suggesting that further research in this area could be highly relevant. This is also supported by Kremsner et al. (2023), who considered it a limitation in FEA software that it is only able to process single frames and no dynamic changes in facial expressions.

Future research could explore incorporating additional biometric data to enhance the correlation between facial expressions and emotional states, potentially leading to more accurate predictions of user experiences. Methods such as EEG or GSR are often seen used in combination with FEA (Santoso et al., 2016; Čertický et al., 2019; Bañuelos-Lozoya et al., 2021; Brunken et al., 2003). Parvanta et al. (2022) state that it can be difficult to estimate the affective state of participants in a non-lab setting without other physiological sensors than facial expression recognition software. However, in the present study, it was not possible to implement EEG or GSR, as these require proper attachment to head and fingers, which the participants would not be able to do themselves. Implementing these would also put the ecological validity at risk, and therefore it might be more beneficial to explore other less invasive biometric approaches, such as screen-based eye-tracking, although this method might provide less informative data compared to EEG and GSR.

Pending the further development and success of a tool for estimating *enjoyment*, *frustration*, *challenge*, *boredom*, and *excitement* in UX contexts through FEA, would be a great asset to UX researchers. One application is to significantly reduce the noise of increased cognitive load by probing participants during think aloud or concurrent probing in user tests. Many researchers refrain from trusting retrospective feedback in user tests, as studies have shown that participants have trouble pinpointing their struggles in retrospect, often either underestimating or overestimating their experience (Abascal et al., 2015). With a tool, as this study presents the first steps towards, UX researchers can pinpoint when participants exhibit experiences of e.g. frustration, and probe them about these events specifically, either in retrospect or concurrently, without having to explicitly pay attention to facial expressions themselves. Further, it allows researchers to quantify the users' experiences without the use of self-assessment scales, that historically contributes a great deal of bias (Bañuelos-Lozoya et al., 2021; The Interaction Design Foundation, 2024).

References

- Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., & Winckler, M. (2015). Mind the gap! comparing retrospective and concurrent ratings of emotion in user experience evaluation. In *Human-computer interaction - interact 2015* (pp. 237–254, Vol. 9296). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-22701-6_17
- Abler, B., Walter, H., & Erk, S. (2005). Neural correlates of frustration. *NeuroReport*, 16(7). <https://doi.org/10.1097/00001756-200505120-00003>
- Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. <https://doi.org/10.1109/WACV.2016.7477553>
- Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. <https://doi.org/10.1109/FG.2018.00019>
- Bañuelos-Lozoya, E., González-Serna, G., González-Franco, N., Fragoso-Díaz, O., & Castro-Sánchez, N. (2021). A systematic review for cognitive state-based qoe/ux evaluation. *Sensors (Basel, Switzerland)*, 21(10), 3439–. <https://doi.org/10.3390/s21103439>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyperparameter optimization. *Journal of Machine Learning Research*, 13(2), 281–305.
- Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(1), 1063–1095. <https://doi.org/10.48550/arXiv.1005.0208>
- Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 53–61. https://doi.org/10.1207/S15326985EP3801_7
- Carmichael, L., Poirier, S.-M., Coursaris, C., Léger, P.-M., Sénécal, S., Davis, F. D., Riedl, R., vom Brocke, J., Léger, P.-M., Randolph, A. B., & Müller-Putz, G. (2021). *Does media richness influence the user experience of chatbots: A pilot study* (Vol. 52). Springer, https://doi.org/10.1007/978-3-030-88900-5_23
- Ceccacci, S., Generosi, A., Giraldi, L., & Mengoni, M. (2023). Emotional valence from facial expression as an experience audit tool: An empirical study in the context of opera performance. *Sensors (Basel, Switzerland)*, 23(5), 2688–. <https://doi.org/10.3390/s23052688>
- Čertický, M., Čertický, M., Sinčák, P., Magyar, G., Vaščák, J., & Cavallo, F. (2019). Psychophysiological indicators for modeling user experience in interactive digital entertainment. *Sensors (Basel, Switzerland)*, 19(5), 989–. <https://doi.org/10.3390/s19050989>
- Chang, D., Yin, Y., Li, Z., Tran, M., & Soleymani, M. (2024). Libreface: An open-source toolkit for deep facial expression analysis. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8190–8200. <https://doi.org/10.48550/arXiv.2308.10713>
- Clark, E. A., Duncan, S. E., Hamilton, L. M., Bell, M. A., Lahne, J., Gallagher, D. L., & O'Keefe, S. F. (2021). Characterizing consumer emotional response to milk packaging guides packaging material selection. *Food quality and preference*, 87, 103984–. <https://doi.org/10.1016/j.foodqual.2020.103984>
- Crist, C., Duncan, S., Arnade, E., Leitch, K., O'Keefe, S., & Gallagher, D. (2018). Automated facial expression analysis for emotional responsivity using an aqueous bitter model. *Food quality and preference*, 68, 349–359. <https://doi.org/10.1016/j.foodqual.2018.04.004>
- Cutler, R. D., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792. <https://doi.org/10.1890/07-0539.1>
- Deng, W., Jia, M., & Zhang, Z. (2023). How corporate social responsibility moderates the relationship between distributive unfairness and organizational revenge: A deontic justice perspective. *Chinese management studies*, 17(6), 1240–1258. <https://doi.org/10.1108/CMS-09-2021-0400>
- de Wijk, R. A., Ushima, S., Ummels, M., Zimmerman, P., Kaneko, D., & Vingerhoeds, M. H. (2021). Reading food experiences from the face: Effects of familiarity and branding of soy sauce on facial expressions

- and video-based rppg heart rate. *Foods*, 10(6), 1345–. <https://doi.org/10.3390/foods10061345>
- Ekman, P., & Friesen, W. V. (1978). Facial action coding system (facs). *APA PsycTests*. <https://doi.org/10.1037/t27734-000>
- Ferrari, R. (2015). Writing narrative style literature reviews. *Medical writing (Leeds)*, 24(4), 230–235. <https://doi.org/10.1179/2047480615Z.000000000329>
- Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of personality and social psychology*, 60(2), 229–240. <https://doi.org/10.1037/0022-3514.60.2.229>
- Gosselin, P., Perron, M., Legault, M., & Campanella, P. (2002). Children's and adults' knowledge of the distinction between enjoyment and nonenjoyment smiles. *Journal of nonverbal behavior*, 26(2), 83–108. <https://doi.org/10.1023/A:1015613504532>
- Gülşen, M., Aydın, B., Güler, G., & Yalçın, S. S. (2023). Ai-assisted emotion analysis during complementary feeding in infants aged 6–11 months. *Computers in biology and medicine*, 166, 107482–107482. <https://doi.org/10.1016/j.combiomed.2023.107482>
- Halamová, J., Kanovský, M., Brockington, G., & Strnádelová, B. (2023). Automated facial expression analysis of participants self-criticising via the two-chair technique: Exploring facial behavioral markers of self-criticism. *Frontiers in psychology*, 14, 1138916–1138916. <https://doi.org/10.3389/fpsyg.2023.1138916>
- Hammond, R. W., Parvanta, C., & Zemen, R. (2022). Caught in the act: Detecting respondent deceit and disinterest in on-line surveys. a case study using facial expression analysis. *Social marketing quarterly*, 28(1), 57–77. <https://doi.org/10.1177/15245004221074403>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
- Holiday, S., Hayes, J. L., Park, H., Lyu, Y., & Zhou, Y. (2023). A multimodal emotion perspective on social media influencer marketing: The effectiveness of influencer emotions, network size, and branding on consumer brand engagement using facial expression and linguistic analysis. *Journal of interactive marketing*, 58(4), 414–439. <https://doi.org/10.1177/10949968231171104>
- Huntington, C. (2024). Frustration: Definition, examples, & principles [Accessed: May 21 2024]. <https://www.berkeleywellbeing.com/frustration.html>
- Jones, T., Randolph, A. B., Sneha, S., Davis, F. D., Riedl, R., vom Brocke, J., Léger, P.-M., Randolph, A. B., & Müller-Putz, G. (2021). *Examining the impact of social video game tournaments on gamers' mental well-being* (Vol. 52). Springer, https://doi.org/10.1007/978-3-030-88900-5_20
- Kassas, B., Palma, M. A., & Porter, M. (2022). Happy to take some risk: Estimating the effect of induced emotions on risk preferences. *Journal of economic psychology*, 91, 102527–. <https://doi.org/10.1016/j.joep.2022.102527>
- Kremsner, T. P., Pfeiffer, C., Weidinger, S., & Stolavetz, C. (2023). How to visualize electricity consumption anomalies: The impact of chart types on triggered emotions and eye movements. *e-Prime*, 5, 100202–. <https://doi.org/10.1016/j.prime.2023.100202>
- Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PloS one*, 13(8), e0199661–e0199661. <https://doi.org/10.1371/journal.pone.0199661>
- López-Mas, L., Claret, A., Bermúdez, A., Llauger, M., & Guerrero, L. (2022). Co-creation with consumers for packaging design validated through implicit and explicit methods: Exploratory effect of visual and textual attributes. *Foods*, 11(9), 1183–. <https://doi.org/10.3390/foods11091183>
- Matsufuji, Y., Ueji, K., & Yamamoto, T. (2023). Predicting perceived hedonic ratings through facial expressions of different drinks. *Foods*, 12(18), 3490–. <https://doi.org/10.3390/foods12183490>
- Mavromoustakos-Blom, P., Kosa, M., Bakkes, S., & Spronck, P. (2021). Correlating facial expressions and subjective player experiences in competitive hearthstone. *ACM International Conference Proceeding Series*. <https://doi.org/10.1145/3472538.3472577>
- Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International journal of remote sensing*, 39(9), 2784–2817. <https://doi.org/10.1080/01431161.2018.1433343>
- McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., & Graesser, A. (2007). Facial features for affective state detection in learning environments. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 29(29). <https://doi.org/10.1145/3385209>
- Mookherjee, S., Lee, J. J., & Sung, B. (2021). Multichannel presence, boon or curse?: A comparison in price, loyalty, regret, and disappointment. *Journal of business research*, 132, 429–440. <https://doi.org/10.1016/j.jbusres.2021.04.041>
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *The American psychologist*, 17(11), 776–783. <https://doi.org/10.1037/h0043424>
- Orne, M. T. (2002). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *Prevention & treatment*, 5(1). <https://doi.org/10.1037/1522-3736.5.1.535a>

- Parvanta, C., Hammond, R., He, W., Zemen, R., Boddupalli, S., Walker, K., Chen, H., & Harner, R. (2022). Face value: Remote facial expression analysis adds predictive power to perceived effectiveness for selecting anti-tobacco psas. *Journal of health communication*, 27(5), 281–291. <https://doi.org/10.1080/10810730.2022.2100016>
- Paul Ekman Group. (2024). Micro expressions [Accessed: May 16 2024]. <https://www.paulekman.com/resources/micro-expressions/>
- Petneházi, G. (2019). Recurrent neural networks for time series forecasting. *arXiv.org*. <https://doi.org/10.1016/j.ijforecast.2020.06.008>
- Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301. <https://doi.org/10.1002/widm.1301>
- Rao, S., Wirjopawiro, S., Pons Rodriguez, G., Röggla, T., Cesar, P., & El Ali, A. (2023). Affective driver-pedestrian interaction: Exploring driver affective responses toward pedestrian crossing actions using camera and physiological sensors. *ACM International Conference Proceeding Series*, 300–310. <https://doi.org/10.1145/3580585.3607168>
- Richlan, F., Thürmer, J. L., Braid, J., Kastner, P., & Leitner, M. C. (2023). Subjective experience, self-efficacy, and motivation of professional football referees during the covid-19 pandemic. *Humanities & social sciences communications*, 10(1), 215–215. <https://doi.org/10.1057/s41599-023-01720-z>
- Rodríguez-Fuertes, A., Alard-Josemaría, J., & Sandubete, J. E. (2022). Measuring the candidates' emotions in political debates based on facial expression recognition techniques. *Frontiers in psychology*, 13, 785453–785453. <https://doi.org/10.3389/fpsyg.2022.785453>
- Samant, S. S., Chapko, M. J., & Seo, H.-S. (2017). Predicting consumer liking and preference based on emotional responses and sensory perception: A study with basic taste solutions. *Food research international*, 100(Pt 1), 325–334. <https://doi.org/10.1016/j.foodres.2017.07.021>
- Samant, S. S., & Seo, H.-S. (2020). Influences of sensory attribute intensity, emotional responses, and non-sensory factors on purchase intent toward mixed-vegetable juice products under informed tasting condition. *Food research international*, 132, 109095–109095. <https://doi.org/10.1016/j.foodres.2020.109095>
- Santoso, H., Schrepp, M., Kartono Isal, R. Y., Yudha Utom, A., & Priyogi, B. (2016). Measuring the user experience. *The journal of educators online*, 13(1). <https://doi.org/10.9743/JEO.2016.1.5>
- Sass, J., & Fekete, L. V. (2022). Secrets revealed by boredom: Detecting and tackling barriers to student engagement. *International Conference on Advanced Learning Technologies (ICALT)*, 417–419. <https://doi.org/10.1109/ICALT55010.2022.00129>
- Savela-Huovinen, U., Toom, A., Knaapila, A., & Muukkonen, H. (2021). Sensory professionals' perspective on the possibilities of using facial expression analysis in sensory and consumer research. *Food science & nutrition*, 9(8), 4254–4265. <https://doi.org/10.1002/fsn3.2393>
- Scikit-learn. (2024a). *Randomforestclassifier* [Accessed: May 27 2024]. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- Scikit-learn. (2024b). *Randomforestregressor* [Accessed: May 27 2024]. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
- Shah, M., Cooper, D. G., Cao, H., Gur, R. C., Nenkova, A., & Verma, R. (2013). Action unit models of facial expression of emotion in the presence of speech. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, 49–54. <https://doi.org/10.1109/ACII.2013.15>
- Sharma, K., Papavaslopoulou, S., & Giannakos, M. (2022). Children's facial expressions during collaborative coding: Objective versus subjective performances. <https://doi.org/10.1016/j.jccci.2022.100536>
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- Tejada, J., Freitag, R. M. K., Pinheiro, B. F. M., Cardoso, P. B., Souza, V. R. A., & Silva, L. S. (2022). Building and validation of a set of facial expression images to detect emotions: A transcultural study. *Psychological research*, 86(6), 1996–2006. <https://doi.org/10.1007/s00426-021-01605-3>
- The Interaction Design Foundation. (2024). Rating scales in ux research: The ultimate guide [Accessed: May 22 2024]. <https://www.interaction-design.org/literature/article/rating-scales-for-ux-research#cons-16>
- Walsh, A. M., Duncan, S. E., Bell, M. A., O'Keefe, S. F., & Gallagher, D. L. (2017). Breakfast meals and emotions: Implicit and explicit assessment of the visual experience. *Journal of sensory studies*, 32(3). <https://doi.org/10.1111/joss.12265>
- Wetzel, E., Bohnke, J. R., & Brown, A. (2016). Response biases. <https://doi.org/10.1093/med:psych/9780199356942.003.0024>
- Zarei, S. A., Yahyavi, S.-S., Salehi, I., Kazemiha, M., Kamali, A.-M., & Nami, M. (2022). Toward reanimating the

- laughter-involved large-scale brain networks to alleviate affective symptoms. *Brain and behavior*, 12(7), e2640–n/a. <https://doi.org/10.1002/brb3.2640>
- Zeng, I. M., & Lobo Marques, J. A. (2023). Neuromarketing as a tool to measure and evaluate the consumer behaviour of guandong teahouse's social media advertisement. *ACM International Conference Proceeding Series*, 63–69. <https://doi.org/10.1145/3616712.3616787>
- Zhi, R., Hu, X., Wang, C., & Liu, S. (2020). Development of a direct mapping model between hedonic rating and facial responses by dynamic facial expression representation. *Food research international*, 137, 109411–109411. <https://doi.org/10.1016/j.foodres.2020.109411>
- Zhi, R., Wan, J., Zhang, D., & Li, W. (2018). Correlation between hedonic liking and facial expression measurement using dynamic affective response representation. *Food research international*, 108, 237–245. <https://doi.org/10.1016/j.foodres.2018.03.042>

Appendix



AALBORG UNIVERSITY

APPENDICIES

Institute for Engineering and Tech
Aalborg University
Engineering psychology
Fredrik Bajers vej 7A
9220 Aalborg Ø

Title:

Applying Facial Expression Analysis Software to Estimate the User Experience of LEGO building by Analyzing Facial Action Units

ECTS:

30 ECTS

Semester:

4. semester, Master's

Semester theme:

Master's Thesis

Project period:

Spring 2024

Project group:

Group nr. 1081

Participants:

Lisa Bondo Andersen
Mathias Thygesen

Supervisor:

Rodrigo Ordoñez

Number of pages: 79**Abstract:**

Denne undersøgelse udforsker brugen af *facial expression analysis* (FEA) til at estimere brugeroplevelsen under LEGO-bygning ved at analysere *facial action units* (AU). Traditionel forskning af brugeroplevelser benytter ofte subjektiv selvrapportering, som kan være underlagt bias. Dette studie forsøger at anvende FEA til at opnå en mere objektiv og kvantificerbar tilgang. I undersøgelsen blev OpenFace 2.0, et værktøj som kan detektere såkaldte *facial landmarks* og genkende AUs, anvendt til FEA. Undersøgelsen involverede et remote setup med 18 deltagere, som byggede et LEGO-sæt derhjemme, mens deres ansigt blev optaget. Disse videoer blev behandlet for at udtrække AU-data, som derefter blev sammenlagt med deltagernes selvrapporterede vurderinger af fornøjelse, frustration, udfordring, kedsomhed og begejstring på en skala fra 1 til 7. På trods af en omfattende datindsamling og analyse resulterede de opstillede *machine-learning*-modeller, baseret på Random Forest regression og klassifikation, en ringe præstation i at forudsige de subjektive vurderinger ud fra AU-data. Dette kan skyldes, at undersøgelsen blev udført i en rum, hvilket kan have reduceret intensiteten i deltagernes ansigtsudtryk. Det kan ligeledes skyldes en utilstrækkelig variation i selvrapporteringer, som følge af at deltagerne ikke blev påvirket tilstrækkeligt i relation til de 5 parametre. Resultaterne antyder, at mens FEA har potentiale for anvendelse i forskning inden for brugervenlighed, er yderligere metodologiske forbedringer nødvendige. Fremtidig forskning bør overveje hyppigere selvrapporteringer, anvende mere komplekse *machine-learning*-modeller og potentielt inkorporere yderligere biometriske målinger for at opnå en forbedring i forudsigelser af emotionelle tilstande.

Preface

The following appendices describe how a remote experiment was conducted, based on a literature review, to collect data on participants' facial expressions and subjective self-assessments, which was used to train machine learning models into predicting the participants' assessments based on their facial expressions.

The study includes an article, describing the main narrative of the study, providing an overview of the research, its objectives, methodology, results, and conclusions. Additionally, you will find complementary appendices and supplementary material. The appendices are detailed documents providing in-depth information, data, and analyses. They describe all aspects of the paper in further detail and can be read both independently or in chronological order. It is, however, recommended to consult an appendix for further information, when met with a reference to a separate appendix to gain the full understanding of the given topic. Supplementary material only refer to non-text files, such as images, videos, data scripts, and CSV files.

In the paper, as well as the appendices, APA is used as the citation style. Up to two authors' last names will appear in the citations along with the date of publication. For the paper, citations will be added directly to the statement, it belongs. For the appendices, the citations will be added after the first statement based on the reference, where the statements that follow will be in relation to the same citation, unless stated otherwise. Citations related to books will have page number present as part of the citation. Figures and tables with no citations are created by the group.

We would like to take this opportunity to give our thanks to our supervisor Rodrigo Ordoñez for his support through the span of this study. A special thanks also goes out to Rasmus Horn, Sr. Quality Manager of the Consumer Perceived Quality department at The LEGO Group, who was our company contact throughout the study, supplying us with LEGO sets for the experiment as well as supporting the entire process of the study.

Contents

	Page
1 Introduction	1
1.1 The LEGO Group	4
1.2 Problem statement	5
2 Literature Review	6
2.1 First round	7
2.2 Second round	9
2.3 Third Round	10
2.4 Self-reported Parameters	12
2.4.1 FEA Software	13
2.4.2 Limitations	13
3 Analyzing Facial Expressions	15
3.1 Action Units and Facial Expression Recognition	15
3.2 Software tools	16
3.2.1 OpenFace functionalities	17
3.3 Measuring emotions with FACS	20
4 Method	22
4.1 LEGO Set	22
4.1.1 Recruitment Survey	24
4.1.2 Experiment Design	27
4.1.3 Pilot Study	33

CONTENTS

4.1.4	Participants	35
4.1.5	Procedure	36
5	Results and Analysis	37
5.1	Data Preparation	37
5.2	Time Spent on Steps	39
5.3	Participant Ratings in Question Sheets	39
5.4	OpenFace Analysis	41
5.4.1	Post-processing of OpenFace Output	42
5.5	Training Models	44
5.5.1	Training Models Utilizing Individual Data Points	47
5.6	Intensity of AUs Compared with Self-reports	50
6	Validation	52
6.1	Mean and Overall Assessment Comparison	52
6.2	Assessment and Debriefing Response Comparison	55
6.2.1	Primacy and Recency	56
7	Discussion	58
7.1	Expression of emotions	59
7.2	Self-assessment Scale	61
7.3	Experimental design	63
7.4	Future work	65
8	Conclusion	68
	Bibliography	70

List of supplementary materials

1. Literature Review
2. Experiment Material
 - 2.1 Instruction Document
 - 2.2 Instruction Video
 - 2.3 Question Sheet
3. Analysis
 - 3.1 Python scripts + data
 - 3.2 ML plots on mean data
 - 3.3 ML plots on individual data

1 Introduction

When measuring the user experience (UX), which refers to a user's experience with an application, product, or service, the goal is to gain understanding and interpret user's perceptions and answers through their perspective (Bañuelos-Lozoya et al., 2021). Traditionally, UX is measured using both quantitative and qualitative methods such as questionnaires and interviews, which typically rely on self-reports and thus are subjective (Santoso et al., 2016; Čertický et al., 2019; Bañuelos-Lozoya et al., 2021). Since subjective data is prone to bias making it less reliable, it is becoming more widely implemented in the methodology throughout the field to gain a more direct and concrete image of the UX by measuring users' cognitive state in terms of e.g. work load, attention and emotions through biometric sensors (Santoso et al., 2016; Čertický et al., 2019; Bañuelos-Lozoya et al., 2021). These biases include demand characteristics, which refers to when participants want to help researchers confirm their hypothesis and, thus, change their behavior accordingly (Orne, 2002). Another example is extreme response style, which is a type of response bias, referring to participants preferring to use the outer points on a scale (Wetzel et al., 2016). The opposite can also be the case, where respondents prefer to use the midpoints on a scale, which is referred to as midpoint response style.

The biometric sensors that are implemented in the methodology include electroencephalogram (EEG), galvanic skin response (GSR), electrocardiogram (ECG), eye tracking (ET) among others, as these methods are more objective and thus do not rely on users to self-report (Santoso et al., 2016; Čertický et al., 2019; Bañuelos-Lozoya et al., 2021; Brunken et al., 2003). EEG is the measurement of electrical activity in the brain and is achieved by attaching electrodes to the scalp (Bañuelos-Lozoya et al., 2021). GSR measures the electrical resistance, determined by an increase or decrease in sweat

production, by attaching electrodes to the middle and index fingers. ECG is measured by placing a set of electrodes on the chest and provides a measurement of electrical activity generated by the heart. ET is typically measured with cameras and by using software that can identify reflection in the cornea and pupil, and establish the related gaze point. These methods have all been used for measurement of emotions, as the body responds differently in all above-mentioned aspects depending on the emotions experienced (Bañuelos-Lozoya et al., 2021). Biometric sensors do, however, often disturb the users' experience with the product they are interacting with, since many of the sensors have to be attached to the user during the interaction, which can limit the movement of certain body parts such as the head or the hands. Facial expression analysis (FEA) is one of the biometric measuring methods that disturb the user experience the least, since no physical sensors have to be attached to them.

FEA is often used to measure emotions (Baltrušaitis et al., 2016; Chang et al., 2024a; Paul Ekman Group, 2024). An *emotion* can be defined as a bodily reaction and cognitive state, involving complex mental processes and behavioral components, typically as a reaction to situations involving the experience of obstacles or progress toward achieving goals (Nordfang and Nørby, 2017, p. 339-342; Ekman, 1992). Basic emotions include happiness, fear, anger, sadness, surprise, and disgust, and are often experienced by an increase in volatile thoughts and arousal, making a person ready to act, which can be expressed both verbally and nonverbally, where the latter is often through facial expressions, gestures or a combination of both. Thus, analyzing facial expressions can be considered an important tool when measuring the UX of products, since they can provide valuable information about users' social and affective states, through a channel of nonverbal communication (Baltrušaitis et al., 2016; Chang et al., 2024a; Paul Ekman Group, 2024). Facial expressions, unlike verbal communication, consists of a universal system of signals reflecting a person's emotional state (Paul Ekman Group, 2024). This means that facial expressions, at least for representing the 6 basic emotions, provides better insight into people's emotions, regardless of culture, language, or personal background compared to verbal and bodily communication. In this regard, micro expressions, which are facial expressions occurring in half a second or less, are important, since they usually happen automatically without the person expressing them even realizing (Paul Ekman Group, 2024). Even if someone tries to conceal an emotion, it will

often still be evident in the face, resulting in a subtle or very quick expression. These expressions can be difficult for other people to detect but give important cues, and could therefore also be an important aspect when assessing the UX of products.

In fields such as UX, data is often based on self-assessments, relying heavily on verbalizing subjective judgments, attitudes, emotional state, and level of stress, among others, which can be influenced by both intrinsic as well as extrinsic factors (Longo, 2018). Apart from this, Longo (2018) also explains that the task of communicating in itself, which for example is done in think aloud testing, can lead to an increase in cognitive load, which affects the outcome of the assessment in a way that it would not in a natural setting of interacting with the product. This unreliability in data collection indicates the need for a data collection method which is not subject to as many biases, and does not require as much effort from the participants, which is why FEA could be an important inclusion in such fields. Even so, we found that in terms of the use of FEA, there is a gap in research since FEA has only been found used to assess UX a limited amount of times and never with the use of tangible products as the stimuli. However, prior research has investigated some aspects of the connection between self-assessed emotions and facial expressions. Here, a correlation has been found between positive and negative emotions in connection with purchase intent in sensory tests (Samant and Seo, 2020). A correlation has also been found between the two for arousal and valence in a nonverbal driver-pedestrian interaction (Rao et al., 2023). It has also been found that facial expressions provide valuable information regarding overall subjective experiences of players in an online video game tournament (Mavromoustakos-Blom et al., 2021). Further on the use of FEA in research can be found in the literature review described in Appendix 2.

When conducting research using FEA, Cohn et al. (2002) found evidence indicating individual differences through unique variance in timing and base rate of how facial expression features were activated, but that these were stable within the individuals through time periods from 4 to 12 months. They state that to make an accurate conclusion regarding emotions, many sources of information such as context and patterns in facial expressions and individual differences among people are required. Moreover, it is also important to consider that at least some facial expressions, such as a smile, might be socially dependant, which means that the presence of other people during data col-

lection of facial expressions might influence the outcome (Fridlund, 1991). However, it is also argued that some features of a smile are automatically produced and, thus, are equally present in solitary and in social scenarios (Schmidt et al., 2003). Through the literature review conducted in this study, we found that only 2 out of 30 reviewed articles researching the connection between self-assessment and FEA carried out experiments with a remote setup (Hammond et al., 2022; Parvanta et al., 2022). These two papers both used videos as stimuli, meaning that no literature was found investigating the relation between FEA and self-assessments using a remote setup and a tangible product as stimuli. This study aims to fill this gap while collaborating with The LEGO group.

1.1 The LEGO Group

The LEGO group was founded in 1932 by Ole Kirk Kristiansen, who at the time had a carpentry workshop where he would make wooden toys (The LEGO Group, 2024b). In 1936, the company got the well-known name "LEGO" which is an abbreviation of the danish words "leg godt" meaning "play well". The LEGO bricks were first produced in 1949 and later redesigned in 1958 to how we know them today. The LEGO Group now produces products that appeal to a wide target audience: LEGO DUPLO building sets for children between 1.5 and 4 years, LEGO System building sets, that use the classic LEGO Bricks, which are mainly for children between 5 and 12 years, and LEGO Technic building sets for older children and teenagers, as well as adults. Apart from these overall categories, the company also produces sets that use a larger amount of LEGO pieces (both classic LEGO bricks and LEGO Technic elements) and more advanced building techniques which are for adults from 18 years and up. Through the years, the company has been passed from father to son, and today it is owned by Kjeld Kirk Kristiansen, who is the grandson of the company founder. The LEGO Group has 37 sales offices, 5 production locations, more than 500 stores and over 24.000 employees around the world. The company's core values are imagination, creativity, fun, care and quality. These values along with Ole Kirk Kristiansen's famous motto "*The best isn't good enough*" is a testament to the company valuing their users and their experiences.

At The LEGO Group, several different departments focus on user experience (UX) at different points in the process, from while the products are being designed to them

being produced and sold to consumers (Internal communication). The company also focuses on UX for many different products, such as their website, their apps, and of course the LEGO sets. Usually, this is done using classical UX methods, such as surveys, personas, or think aloud user tests. This is done both before and after releasing products. One of the departments focusing on feedback from users is *Consumer Perceived Quality* (CPQ), which is the department we collaborated with as part of this study. Eight people work in this department, focusing on the website, apps, and physical products. They focus on customer feedback received after products have been released. In CPQ, the employees are always looking for new ways to collect data on UX to be used and analyzed in various ways. This is also why they were very interested in the idea of biometric data being used in the form of FEA, since this study could lead to new ways of using this method for data collection.

1.2 Problem statement

The aim of this study was to develop a method to collect data in a remote setting that could be used for FEA to evaluate UX, which could later be prepared in a way to be used in machine learning models to predict relevant UX metrics. For this, the aspects of conducting research using FEA were considered, which ended up leading to the following problem statement:

How can a remote experiment be conducted to collect data for FEA to evaluate UX metrics while building with LEGO?

The goal of collecting and analyzing the facial expressions is to research the correlation between facial expressions and UX metrics in the context of building LEGO. Thus, the facial expressions would function as an estimator for these metrics as an addition to, or potentially a replacement of, verbal feedback. To reach this goal, a literature review was first conducted to investigate the uses of FEA in prior research (See Appendix [2](#)). Then, an experiment was carried out using a remote setup and a LEGO set as stimuli to investigate the connection between participants' self-assessments and their facial expressions (See Appendix [4](#)).

2 Literature Review

As the initial part of the study, a literature review was carried out. The purpose of this was to review how prior research has used facial expression analysis (FEA) and how this could be used as inspiration for carrying out our experiment. Another purpose was to research whether or not FEA had been used in UX related contexts, e.g. in order to estimate UX metrics of participants interacting with products. This was done in order to discover if and where there was a gap in research that this study could fill. Snyder (2019) and Ferrari (2015) describe three different literature review approaches; *systematic*, *narrative*, and *integrative* review. As the purpose of our review mainly was to gain inspiration for the experiment later carried out, we decided on conducting a narrative review. A systematic review is typically used to compare results and measure effect sizes, and it has an unnecessarily rigorous structure for the purpose of this review. The narrative review, however, fits this case because it allows the researcher to look across contexts and groups within the same topic, and thus would create a larger sample in a field with relatively few publications. While the integrative review is closely related to the narrative review, its aim is to construct new theoretical frameworks and perspectives through assessing, critiquing, and synthesizing the literature on a research topic, which was also not the purpose of the review.

The structure of the review was inspired by Snyder (2019) and Ferrari (2015). Table 2.1 shows an overview of the three rounds of the literature review described in this appendix. The review consisted of three rounds in total. For the first and second round, specific search prompts were used to find papers on two priorly selected databases to determine whether they should be included or rejected in the review. The inclusion decision was based solely on reading through the abstract of the papers. In the third and final round, the papers from the second round were read all the way through to further

establish the method, parameters, limitations and so on for the included papers. Further details on all articles reviewed in the literature review can be seen in supplementary material 1.

Table 2.1: The table shows the inclusion and rejection criteria used in all three rounds of the literature review and the number of abstracts/papers read and included in each round. For the first and second round, only abstracts were reviewed, while the entire article was read through for the last round.

Round	Inclusion Criteria	Rejection criteria	Read	Included
1st	1. Measures facial movement in relation to a psychological response	1. Review articles 2. Development of FEA	99	59
2nd	1. Exposure to stimuli 2. Measures facial movement related to mood, emotion etc. compared with self-reported subjective assessment	1. Medical research unrelated to emotions	202	33
3rd	1. Included papers from the 2nd round	1. Not written in English/Danish 2. No access	33	30

Before carrying out the first round of the review, the databases to use were defined. Here, APA PsycInfo and Scopus were chosen, because these databases contain literature within the desired field. Only articles and conference papers were included in the review, which means that e.g book chapters and user manuals on how to conduct experiments using FEA were not included. This is again due to the fact that we wanted to gain inspiration from methods, and book chapters and user manuals do not include this. After defining the databases and the type of material to include, the first round of the review was conducted.

2.1 First round

To assess how broad the field was, the following search prompt was used in the selected databases: “*Facial expression analysis*”. The inclusion criteria was *literature that measures facial movement in relation to a psychological response*. This means that e.g. review articles and articles that aimed to develop a new analysis method or algorithm using FEA were rejected, because no experiment was conducted. Besides this, the inclusion criteria meant that a lot of papers would be included, which was desired since the

purpose of this round was to be explorative and include papers in various different areas of the field of FEA. To determine whether the papers should be included in the review, we read through the abstract. Since the purpose of this round was to get an overview of the field, we decided to not read the full article yet as the relevant information was accessible from the abstract.

While reading the abstract, the following information was noted down: *One author's last name, publication year, doi, article name and the database it was found in.* Aside from these, *physiological measure(s) besides FEA and sensors* were noted down to determine whether or not FEA often was used in connection with other physiological measures, such as eye-tracking or EEG and if so, how this was done, which could be used as inspiration for the experiment in this study or as part of future work. *The psychological parameter*, such as specific emotions, engagement, or pain and *tool to measure psychological parameter*, which could be specific types of surveys or interview questions, was noted down as well. The last aspect noted down was the *stimulus*, i.e. what the participant interacted with/was shown during the experiment, such as a video game or an ad video. This was used to determine if the context was relevant for our study.

In the first round, a total of 99 abstracts were reviewed (65 on APA PsycInfo and 34 on Scopus), which resulted in the inclusion of 59 articles/conference papers. We did not review all articles and conference papers in the databases that appeared from the search prompt. Initially, we had defined a stop criteria after *10 consecutive rejected results* as recommended by Snyder (2019) and Ferrari (2015), but we ended the round prematurely as the inclusion criteria were defined too loosely. The first round provided insights that lead us to narrow down the search prompts and criteria further for the second round. For instance, the first round included literature with experiments involving FEA using animals as subjects. Furthermore, a large part of the reviewed literature included clinical trials, where FEA was either used on subjects with different diagnoses, e.g. to improve the diagnostic process, or to detect pain levels from patients. Because of the defined inclusion criteria, the papers of the above-mentioned fields were also included in the first round of the review. It was not ideal to include those papers in the final round of the review, which is why another round was conducted.

2.2 Second round

For the second round, we decided to focus more specifically on the same area as the experiment in this study, by only including articles and conference papers that involved subjecting them to some sort of stimuli as part of the experiment. This stimuli could either be a tangible product they interacted with, a service, or a video/ad they were watching. Other papers researched e.g. neurodivergent participants and their facial expressions but without subjecting to a stimuli, and were therefore not included in this round. Furthermore, a rejection criteria was implemented ensuring that e.g. medical research such as that described in the first round would not be included again. Another criteria was *measuring facial movement response related to e.g. an emotion compared with a subjective assessment*. It is worth noting that this implicates the inclusion of human participants, since at least none of the reviewed studies specify self-assessment done by animals, which eliminates the problem of having to include research with non-human subjects. Furthermore the rejection criteria from the first round were also used in the second round to ensure that review articles and papers with an analysis or algorithm development purpose were rejected. The narrowing of the field area is also reflected in the search prompt, which was as follows: “*“facial expression analysis” AND (“scale” OR “survey” OR “interview”) NOT “pain”*”. This way, it would increase the likelihood of studies using self-assessment from participants to appear, and for medical research to be less likely to appear, since these often were related to pain.

In the second round, a total of 202 abstracts were reviewed (81 on APA PsycInfo and 121 on Scopus), which resulted in the inclusion of 33 papers. The abstracts reviewed in this round also involved the papers included from the first round. All abstracts from literature found via APA PsycInfo that matched the search prompts were reviewed, while on Scopus, 10 consecutive papers were rejected, leading to the round ending. To ensure the inclusion of as much relevant literature as possible, the criteria from the second round of the review was repeated replacing *“facial expression analysis”* in the prompt with *“facial expression recognition”* as well as *“facial action units”*. This gave both new results and results that also showed using the old prompt, but no new papers were included with the new prompts.

2.3 Third Round

The purpose of this round was to assess the relevant papers through a deeper analysis of the methods, software tools, stimuli, etc. All papers that were included in the second round were read in this round of the review. However, three of the papers from the second round were not included, as they were either not available through the university database or were not written in English or Danish.

Through the review, it was found that FEA is used broadly in different contexts and with different kinds of products/stimuli. The most common contexts found in the review was while watching videos on computer screens, e.g. ads, public service announcements, emotional videos, video lectures, and Instagram posts (Walsh et al., 2017; Holiday et al., 2023; Zeng and Lobo Marques, 2023; Parvanta et al., 2022; Hammond et al., 2022; Kassas et al., 2022; Zarei et al., 2022; Rodríguez-Fuertes et al., 2022; Sass and Fekete, 2022). Parvanta et al. (2022) distributed a survey with three 30-second videos related to an *anti-smoking* campaign and multiple choice questions. Here, the participants could complete the survey at their own pace, while their face was recorded when watching the videos. Through the survey, the participants assessed the effectiveness of the service announcements, their own desire to stop smoking and to share the videos on social media platforms, as well as their willingness to complete a *Quit Now* trial.

Comparably, FEA has also been seen used in other contexts on PCs, such as chatbot interactions, video game tournaments, and online shopping experiences (Jones et al., 2021; Mavromoustakos-Blom et al., 2021; Carmichael et al., 2021; Mookherjee et al., 2021). Mavromoustakos-Blom et al. (2021) conducted an experiment using video game tournaments as stimuli, where a total of 17 players participated, resulting in a total of 31 matches being played. The experiment was carried out in a lab setting using five pairs of high-end gaming desktop computers placed in two rows facing each other. In each match, the opposing players were placed facing each other and the monitors were set to minimum height to enable eye contact between the players. The players' faces were recorded during the matches with webcams mounted on top of each monitor.

Another context FEA is used in is sensory tests, using odor, food, or drink samples as stimuli (Gülşen et al., 2023; Matsufuji et al., 2023; Crist et al., 2018; Savela-Huovinen

et al., [2021]; de Wijk et al., [2021]; Zhi et al., [2020]; Samant and Seo, [2020]; Samant et al., [2017]; Zhi et al., [2018]). An example of this is a study by Samant and Seo ([2020]), who asked participants to evaluate five solutions of mixed-vegetable juice products. This included an image of the packaging, with the price displayed beneath, as well as smelling and tasting the sample. During the entire test, physiological responses including facial expressions, heart rate, skin temperature, and skin conductance response were measured, and later compared to self-reported assessments of overall liking, aroma, flavor (sweetness, sourness, bitterness, and saltiness), and purchase intent. The physiological responses were measured for 15 seconds before the test started, in order to get an individual baseline. Next, participants evaluated the sample first by only looking at an image, followed by being asked to smell the sample to evaluate the aroma. Finally, they evaluated the sample by tasting it.

FEA has been seen less commonly used in a physical experience context using a football game, a driver/pedestrian interaction, or an opera performance as stimuli (Richlan et al., [2023]; Rao et al., [2023]; Ceccacci et al., [2023]). It was also found that FEA was used to measure facial expressions during self-critique, looking at different chart types indicating energy consumption, and in a corporate social responsibility scenario (Halámová et al., [2023]; Kremsner et al., [2023]; Deng et al., [2023]).

Lastly, FEA has been seen used with tangible products, e.g. in the case of packaging assessments (Clark et al., [2021]; López-Mas et al., [2022]). Clark et al. ([2021]) measured product-associated emotions using a check-all-that-apply sheet and an association test, where the participants e.g. were presented to images of two milk packaging samples with a neutral or emotional word between them. The participants were then asked to choose between the packaging samples to categorize them using the word. During the test, the participants also interacted with the packaging samples as if they were inspecting them in a grocery store. Product-associated emotions, product acceptability and purchase intent was measured after the interaction and later compared with the participants' facial expressions during the test. This means that no papers were found where they focused on the entire product experience, from the packaging to interacting with the product itself. However, outside of the literature review, an article by Roy et al. ([2020]) was found. They conducted an experiment assessing the user experience of unboxing, installing and using a TV, while participants were wearing electroencephalo-

gram (EEG) and galvanic skin response (GSR) sensors, and we see no reason why this can not be replicated using FEA with LEGO building, as this context can be considered comparable in or perhaps even lower in complexity than a TV installation.

2.4 Self-reported Parameters

Regarding the self-reported subjective assessments, the most commonly used ones were the 6 basic emotions (happiness, fear, anger, sadness, surprise, and disgust), as well as arousal and valence (Richlan et al., 2023; Deng et al., 2023; Clark et al., 2021; Walsh et al., 2017; Gülşen et al., 2023; Rao et al., 2023; Matsufuji et al., 2023; Kremsner et al., 2023; Zeng and Lobo Marques, 2023; Crist et al., 2018; Ceccacci et al., 2023; Kassas et al., 2022; Zarei et al., 2022; Rodríguez-Fuertes et al., 2022; Sass and Fekete, 2022; Mavromoustakos-Blom et al., 2021; Savela-Huovinen et al., 2021; de Wijk et al., 2021; Jones et al., 2021; Carmichael et al., 2021; Samant et al., 2017).

The parameters engagement, desire to stop smoking, attention, self-criticism, regret and disappointment, purchase intent, and liking were measured by one paper each (Holiday et al., 2023; Parvanta et al., 2022; Hammond et al., 2022; Mookherjee et al., 2021; Halamová et al., 2023; López-Mas et al., 2022; Zhi et al., 2018; Zhi et al., 2020). Rodríguez-Fuertes et al. (2022) combined measuring emotions with measurements of engagement, while de Wijk et al. (2021) and Samant et al. (2017) combined measuring emotions and liking.

Most papers collected the self-reported parameters through a scale presented in a survey/questionnaire at the end of the experiment (Richlan et al., 2023; Clark et al., 2021; Walsh et al., 2017 etc.). Others collected the self-reports in an interview (Kremsner et al., 2023), or by presenting a scale to the participants one or multiple times during the experiment (Rao et al., 2023; Matsufuji et al., 2023; Savela-Huovinen et al., 2021 etc.). Some of the reviewed papers conducted an interview at the end of their experiment in addition to ranking one or more parameters, usually to validate data retrospectively (Richlan et al., 2023; Rao et al., 2023; Sass and Fekete, 2022 etc.). Finally, another group of papers, did not ask participants to rank emotions, but rather to choose between which specific emotions described their current state the best or to indicate whether they were feeling negative or positive emotions (Kassas et al., 2022; Jones et al., 2021; Clark

et al., 2021).

2.4.1 FEA Software

During the review, several of the studies used the same software tools. An overview of which tools were used and in how many articles can be seen in Table 2.2. The table shows that FaceReader was the most commonly used software among the considered papers in the review (Richlan et al., 2023; Deng et al., 2023; Clark et al., 2021 etc.), where the second most commonly used software was iMotions (Kremsner et al., 2023; Parvanta et al., 2022; Hammond et al., 2022 etc.). These two software platforms require a paid subscription to be used, whereas OpenFace, which was used by two papers, is an open source software tool (Gülşen et al., 2023; Mavromoustakos-Blom et al., 2021). Both Affectiva AFFDEX and the FDOF algorithm were used by one paper each (Kassas et al., 2022; Zhi et al., 2020). iMotions uses the AFFDEX technology as well, but offers a GUI for easier interaction with the data, and it is unclear from the method described by Kassas et al. how they employed the technology.

Table 2.2: The table shows the different FEA software tools used by the papers in the final round of the literature review. The table also shows how many papers used each software.

Software used	FaceReader	iMotions	OpenFace	AFFDEX	FDOF	Software not described
No. of papers	12	6	2	1	1	2

2.4.2 Limitations

During the review, we researched which limitations the authors of the papers considered after conducting their experiment. One of these limitations was described by Walsh et al. (2017) who had to exclude participants with excessive facial hair or glasses to get the best results from the FaceReader software. Gülşen et al. (2023) described multiple limitations, one of which being that the confidence score the OpenFace software generates for each frame it examines, decreases significantly when participants turn away from the camera. They also state that the OpenFace software encounters challenges in differentiating between AU10 (upper lip raiser) and AU12 (lip corner puller) because of their similarity. Kremsner et al. (2023) point out that only single images are evaluated

in facial expression analysis, which means that dynamic changes in facial expression are not evaluated.

Parvanta et al. (2022) state as a limitation that it is difficult to estimate the affective state of subjects in a non-lab setting without other physiological sensors than FEA software. This is assumed to be due to the fact that many unknown factors might affect the participants' mood during the experiment, e.g. from objects in their surroundings. They also state that the stimuli, which in this case was public service announcements regarding tobacco, might trigger emotions from past experiences of the participants. This is also described by Clark et al. (2021), who used milk packaging as stimuli, but found that past experiences could potentially influence the participants' assessment during the experiment, because they chose common packaging for a very common product. However, this bias was limited by the fact that the experiment had a within-subject design. Sass and Fekete (2022) conducted an experiment where students were asked to watch a video lecture while their face was recorded. Here, they describe as a limitation that self-reported emotions have only very limited suitability, especially with regards to the detection of boredom. Furthermore, they state that the participants might be influenced by the observational situation, which could have reduced the level of boredom appearing on their face.

3 Analyzing Facial Expressions

Automated software has been used widely to detect the presence of faces as well as the facial features by analyzing the behavior of the face (Chang et al., 2024a). Facial expression analysis (FEA) refers to the automated process of detecting faces and their expressions, often done with machine learning algorithms. Chang et al. (2024a) distinguish between two methods that are most commonly used within FEA; *facial action unit intensity estimation* and *facial expression recognition*.

3.1 Action Units and Facial Expression Recognition

Facial action units (AU) refer to how either individual or groups of facial muscles are activated. AUs were first formalized by Ekman and Friesen (1978) in the Facial Action Coding System (FACS). FACS is typically used for measurement of emotion, and is distributed into 44 distinct AUs as well as multiple categories of head poses and eye gazes, where each AU is enumerated. The system was created to include all distinguishable facial movements into distinct categories, not only related to emotion-specific movements, but rather focused on any anatomically possible facial movements (Ekman and Rosenberg, 2005, p. 13-15). This means that for e.g. *Lip Pressor* (AU24), multiple muscles are activated, but these will in almost every case be activated simultaneously and are therefore categorized into the same AU. Thus, there is not a direct correspondence between single muscles and AUs, and furthermore, some muscles can act in different ways and activate different AUs depending on how they are activated. This is seen for

e.g. *Inner Brow Raiser* (AU1) and *Outer Brow Raiser* (AU2) which are activated by the same muscle. Each AU is designed to be recognized as either present (indicated by 1) or not present (indicated by 0), and if present, an intensity can be determined ranging from 1-5 (Ekman and Rosenberg, 2005, p. 15). Since the formalization of FACS, people have been certified as coders to learn how to identify the presence and intensity of AUs, and in later years, such coded data sets have been used to train software to recognize the AUs (Ekman and Rosenberg, 2005, p. 13).

Facial expression recognition (FER) refers to identifying specific expressions, or rather emotions by identifying the prototypical facial displays (Chang et al., 2024a). This is a complex task because of how much facial expressions can vary between different people.

3.2 Software tools

Various different toolkits have been developed to detect and recognize AUs and facial expressions. Table 3.1 shows a selection of these toolkits. The table shows that all tools offer *Landmark* detection, a technique used to identify facial features (face shape, mouth, nose, eyes etc.), as well as *real-time* analysis of videos. However, AFFDEX and FACET only offer paid services, and dlib, FaceTracker, and Mediapipe do not offer *AU* intensity estimation. *Train* and *Test* refer to whether users can use the tool's algorithm to train and test using their own data sets, i.e. whether the code for inference is released.

Table 3.1: Comparison of various FEA tools showing their features and capabilities. The table was created with inspiration from Chang et al. (2024a). AU refers to AU intensity estimation and FER stands for facial expression recognition.

Tool	Landmark	AU	FER	Train	Test	Real-time	Free
dlib	✓			✓	✓	✓	✓
FaceTracker	✓				✓	✓	✓
Mediapipe	✓		✓		✓	✓	✓
AFFDEX 2.0	✓	✓	✓			✓	
FACET	✓	✓				✓	
OpenFace 2.0	✓	✓		✓	✓	✓	✓
LibreFace	✓	✓	✓	✓	✓	✓	✓

Chang et al. (2024a) describes that as of 2024a, *OpenFace*, developed by Baltrušaitis et al. (2016) and later refined to OpenFace 2.0 by Baltrušaitis et al. (2018), is the most commonly used free toolkit, and it offers all the features shown in Table 3.1 except for FER.

Several studies have shown promising results using deep learning based methods for FER resulting in more robust estimations compared to formerly released tools using more classical machine learning models such as *Support Vector Machine* (SVM) for recognition of AUs (Chang et al., 2024a; Wong et al., 2020). However, these models often require stronger computational power and require larger training data sets. For this reason, running real-time analyses have previously been difficult. However, Chang et al. (2024a) has created their tool, *LibreFace*, in an attempt to combine the functionality of detecting AUs (with a similar or better approximation than OpenFace and alike tools), and by including the more complex functionality offered by FER, while still maintaining a relatively low computing time.

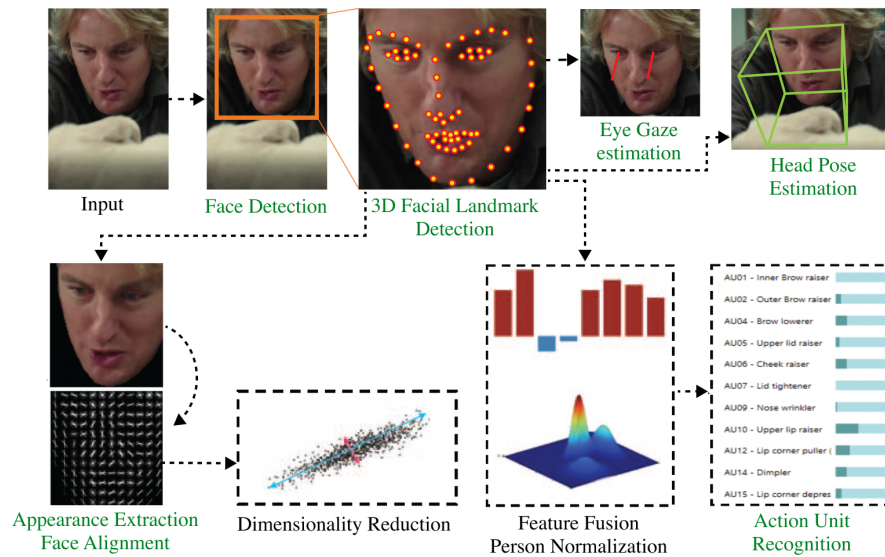
For these reasons, we initially attempted to use the LibreFace tool to analyze the videos gathered in the experiment described in appendix 4. However, due to limitations in access to computers with the proper processing power, since LibreFace only runs on Windows for video analysis, and after multiple attempts with our available resources resulting in outputs with missing or errors in data, we decided to go in a different direction (Chang et al., 2024b). Instead, we managed to run the videos through the OpenFace software with better and more consistent results, although without the FER approach that LibreFace offers. However, Baltrušaitis et al. (2018) argue that their AU estimation in OpenFace was as competitive as the developed deep learning methods at the time of their publication.

3.2.1 OpenFace functionalities

OpenFace 2.0 offers functionalities such as facial landmark detection, head pose estimation, eye gaze tracking, and AU recognition (Baltrušaitis et al., 2018). The process can be seen in Figure 3.1.

An input of an image, either derived from a video sampling with about 30-40Hz or from a single image upload, is received by the model, in which the tool will detect whether a face is present or not (Baltrušaitis et al., 2018). Facial landmark detection

Figure 3.1: The figure shows the steps of how each frame gets processed by OpenFace 2.0. It is based on graphics from the research done by Baltrušaitis et al. (2018).





















then identifies key points on the face, such as the corners of the eyes, the tip of the nose, and the edges of the lips. The system works by detecting the overall face first and then refining the position of each landmark through a series of iterations, to increase precision. Head pose estimation is determined by using the detected facial landmarks to build a 3D representation of the face, and the system then calculates the head's orientation by solving geometric problems related to the position of these points. OpenFace estimates eye gaze by first detecting the positions of the eyelids, irises, and pupils. It then uses these points to calculate the direction of the gaze. The system projects a ray from the camera through the pupil's center and intersects it with a model of the eye to determine the gaze vector.

OpenFace can detect both the presence and intensity of various AUs (Baltrušaitis et al., 2016; Baltrušaitis et al., 2018). Table 3.2 shows the AUs that OpenFace 2.0 can detect. The detection process involves several steps:

1. **Appearance Extraction:** The system extracts features from the face using *histograms of oriented gradients* (HOGs) and shape features from the detected landmarks. These features effectively represent the facial appearance and structure.
2. **Dimensionality Reduction:** The high-dimensional data is then reduced to be han-

Table 3.2: The table shows the AUs OpenFace 2.0 detects (Baltrušaitis et al., 2018). Predictions are available for intensity and presence for all AUs, except for AU28 where only presence is predicted.

AU	Full name	Illustration
AU1	Inner brow raiser	
AU2	Outer brow raiser	
AU4	Brow lowerer	
AU5	Upper lid raiser	
AU6	Cheek raiser	
AU7	Lid tightener	
AU9	Nose wrinkler	
AU10	Upper lip raiser	
AU12	Lip corner puller	
AU14	Dimpler	
AU15	Lip corner depressor	
AU17	Chin raiser	
AU20	Lip stretched	
AU23	Lip tightener	
AU25	Lips part	
AU26	Jaw drop	
AU28	Lip suck	
AU45	Blink	

dled efficiently by using principal component analysis to simplify data structures.

3. **Support Vector Machines (SVM):** SVMs are used to classify the AUs. The training of these models are based on a number of large validated facial expression datasets (DISFA, CK+ etc.)
4. **Person-specific Normalization:** To account for individual differences in facial expressions, the system normalizes the features for each person. This involves adjusting the predictions based on the median values of the features for each in-

dividual to reduce bias and enhance the model’s accuracy.

5. **Prediction:** Once trained, the SVMs can predict both the presence and the intensity in new facial images. The system includes a correction mechanism for AU intensity predictions. It adjusts the predictions by subtracting the lowest n_{th} percentile of the predictions for a specific person, reducing bias and improving consistency.

3.3 Measuring emotions with FACS

FACS is typically used to measure facial expressions of emotions, and certain AUs are known to be active with certain emotions (Ekman and Rosenberg, 2005, p. 17). Research shows general agreement of which AUs are active for the basic emotions; happiness, sadness, anger, fear, surprise and disgust (Ekman and Friesen, 1978; Tejada et al., 2022; Sharma et al., 2022). However, the five parameters, enjoyment, frustration, challenge, boredom, and excitement, which the participants were instructed to self-assess in the experiment (described in appendix 4) are less researched in literature. Table 3.3 shows which AUs are assumed to be active when each parameter is rated highly, and which emotions have been shown to result in similar expressions to the aforementioned basic emotions.

Table 3.3: The table shows the 5 parameters used in the experiment, and which AUs these activate as well as which AUs are activated by similar emotions. Happiness* indicates that the AUs active for this emotion are based on assumptions and not researched in any papers.

Parameter	Similar emotion	AU active	Reference
Enjoyment		AU6, AU12	Perron and Roy-Charland, 2013; Gosselin et al., 2002
	Happiness	AU6, AU12	Tejada et al., 2022; Ekman and Friesen, 1978
Frustration		AU12, AU43	Sharma et al., 2022; McDaniel et al., 2007
	Anger	AU4, AU5, AU7, AU23	Tejada et al., 2022; Ekman and Friesen, 1978
	Sadness	AU1, AU4, AU15	Tejada et al., 2022; Ekman and Friesen, 1978
Challenge		-	
Boredom		AU4, AU7, AU12	Sharma et al., 2022
	Neutral	none	McDaniel et al., 2007
Excitement		-	
	Happiness*	AU6, AU12	Tejada et al., 2022; Ekman and Friesen, 1978

For experiences of *enjoyment*, we see that AU6 and AU12 should be active, and that this is also the case when experiencing happiness, which according to Gosselin et al. (2002)

results in similar expressions as feelings of enjoyment. We did not find any literature on facial expressions during feelings of *excitement*, but we made the assumption that happiness would also correlate with this emotion, and so AU6 and AU12 are assumed to be active when feeling excited. For feelings of *frustration*, AU12 and AU43 were shown to be active by Sharma et al. (2022) and McDaniel et al. (2007). However, OpenFace does not measure AU43, which is eye closure, so to not confuse frustration with enjoyment, we could also look into the AUs activated by anger and sadness, as frustration has been reported to include feelings of anger and sadness (Huntington, 2024; Abler et al., 2005). No AUs have been found in literature to correlate with feelings of *challenge*, and neither have any of the basic emotions. *boredom* was found by Sharma et al. (2022) to activate AU4, AU7 and AU12 and by McDaniel et al. (2007) to activate no AUs, similar to that of a neutral facial expression. Furthermore, Sass and Fekete (2022) describe that facial expressions of boredom are easily influenced by being observed, and might be concealed in such a situation.

4 Method

The purpose of this experiment was to investigate whether a correlation could be found between participants' facial movements (actions units) and their self-reported subjective assessments of 5 parameters; *enjoyment*, *frustration*, *challenge*, *boredom*, and *excitement*. Further, if an experiment procedure using a remote setup can be conducted for collecting useful data to be used in facial expression analysis (FEA). Following a successful data collection, next steps involve training machine learning models on the action units to predict the self-assessment parameters.

The experiment was carried out as a remote study, where the required materials were delivered to the participants so they could complete it at home at their own pace. This was chosen as people typically build with LEGO at home, leading to the assumption that the ecological validity of the experiment would be increased (Orne, 1962). This is also supported by Sass and Fekete (2022) who state that participants' facial expressions are likely to be influenced by being observed. Furthermore, we decided to not tell the participants, we were going to analyze their facial expressions as it was assumed that this could have also affected their expressions because of demand characteristics, where the participants would want to help us confirm our hypothesis (Orne, 2002). Instead, the participants were initially just told that they had to build a LEGO set while recording themselves.

4.1 LEGO Set

Through collaboration with our company contact, we decided that the product the participants were going to interact with during the experiment would be the LEGO Set 10313 called *Wildflower Bouquet* as seen on Figure 4.1. This set is one of the most

Figure 4.1: The figure shows the *Wildflower Bouquet* LEGO set that the participants built as part of the experiment. The figure is from The LEGO Group (2024c).



popular LEGO sets of 2023/2024 and it is intended for the age group 18+ (The LEGO Group, 2024c; internal communication). The CPQ team mentioned that this set has a wide target group spanning across a large age group as well as between genders. Using this set in the experiment would result in the ability to recruit participants with great variety in appearance, leading to a stronger data set and analysis. The set includes 939 pieces, distributed in enumerated bags from 1-4. It contains eight different species of wildflowers with a total of 17 flowers, as some flowers are built more than once.

Initially, the experiment was supposed to include LEGO set 71360 called *Adventures with Mario Starter Course*, because this set focuses on play in a different way than traditional LEGO sets (The LEGO Group, 2024a). Moreover, our company contact at LEGO had observed some issues with how the users interacted with the set that could be interesting for us to look into. From his point of view, the issues arose because the user has to interact with an app while they are building the set. The app includes the user

manual for building as well as inspiration for play during and after finishing building. This switch between app and real-world had shown to be challenging for the young age group the set was designed for, which is children 6 years and up.

However, after interacting with the set, we decided to go in a different direction for several reasons. First of all, remote testing could become difficult with children at the age of 6. Even if we somehow managed to create a proper experiment setup with children, it would most likely not be generalizable to adults since children tend show facial expressions in a different way than adults (Lawrence et al., 2015). A solution to this could have been to conduct the experiment with a different target group than the intended one, but we did not think the set would allow for enough challenge for adults for their ratings in e.g. challenge and frustration to be above a neutral rating leading to the data potentially being severely skewed. Secondly, results from a wrong target group would not be very helpful to the CPQ team at LEGO.

4.1.1 Recruitment Survey

Participants were recruited by distributing a survey through our circle of acquaintances. The flow of the survey can be seen on figure 4.2. On the first page, the respondents were thanked for their consideration to participate in the experiment, and then given an introduction to the experiment. Furthermore, the first page contained a consent form explaining what kind of information would be collected and that it would be stored according to GDPR rules until July 2024, after which it would be deleted. It was also stated that they could ask questions or withdraw their consent at all times by contacting the researchers. The respondents gave consent by clicking *next* in the survey. On page two, first name, gender, age, highest completed level of education, and occupation were collected. The page also had a section where the respondents were asked what device they would record the video of them completing the experiment with, after which they were informed that they should make sure that the device had at least 10 GB of space free to be able to record the entire experiment. Lastly, the page presented the question *"Do you wear glasses to see or read?"*.

On the third page, the respondents were presented with different questions depending on prior answers. This page was mainly focused on whether or not they had facial hair and the respondents' use of glasses, as this was found to be a problem when using

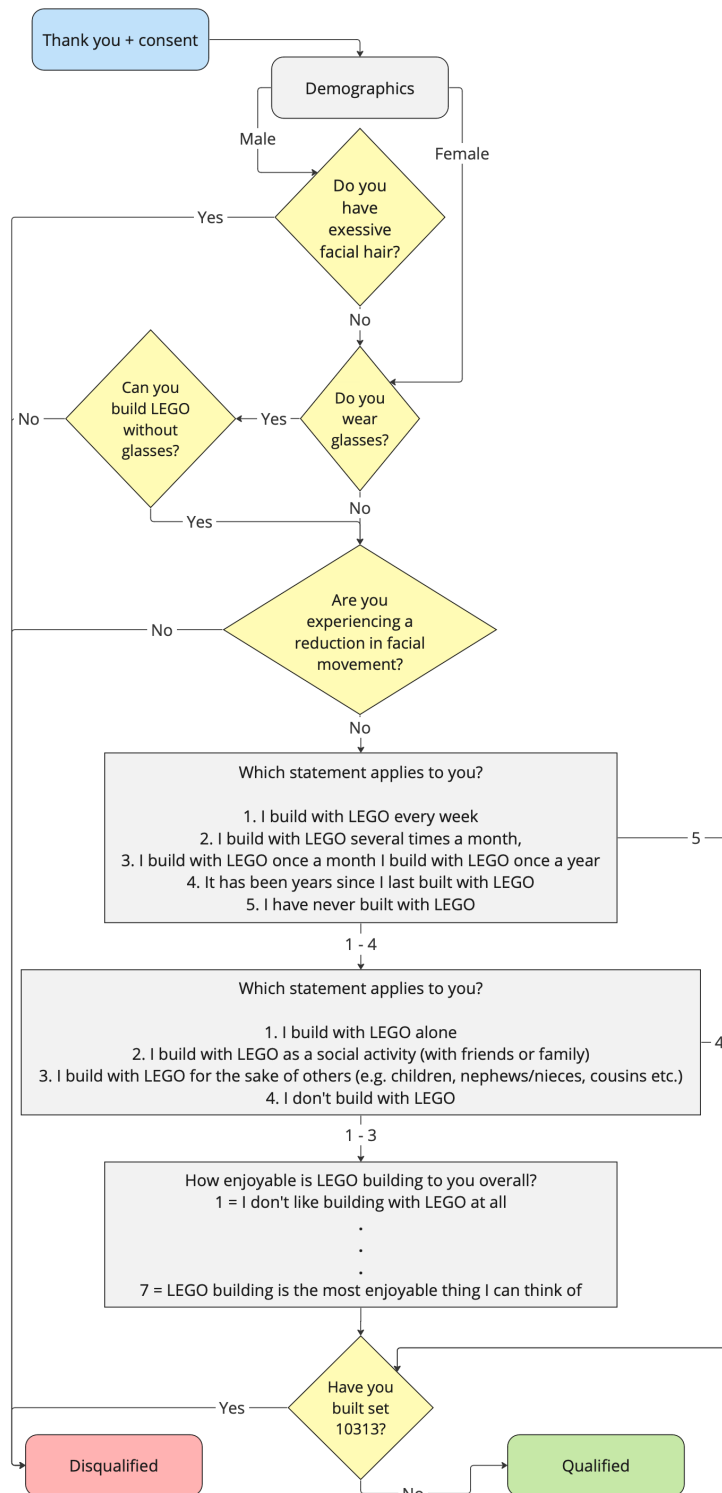
facial recognition software (Clark et al., 2021; Crist et al., 2018). Lastly, regardless of prior answers, they were asked the following *"Are you experiencing a reduction in facial movement? (e.g. due to an illness or injury)"* as this could affect how they express facial expressions.

The fourth page was meant to establish the respondents' relation to LEGO building, both in general and regarding set 10313. The questions on this page were based on Zel et al. (2021), who conducted an experiment about learning, and for this reason asked participants about their motivation to learn the topic as well as external and environmental factors. In our case, the respondents were asked how often they build with LEGO ranging from every week to never having built with LEGO. Next, they were asked if they usually built with LEGO alone, with friends and family or for the sake of others. The respondents were also asked to assess how enjoyable LEGO building was to them on a scale from 1-7. Lastly, the respondents were asked if they had built LEGO set 10313 before, while also being presented with a picture of the set.

If the respondents answered that they did not have facial hair that hid their jawline, that they could complete the experiment without the use of glasses, that they did not experience a reduction in facial movement, and that they had not built LEGO set 10313 before, they were presented with a page telling them that they were qualified to participate in the experiment. Next, they were asked to enter their e-mail, making it possible to contact them regarding the participation. If on the other hand, the respondents did not qualify for the experiment, they were presented with a disqualification page thanking them for their time and telling them that they unfortunately did not meet the criteria to participate in the experiment. The survey was designed so that the flow would end, leading them directly to the disqualification page, if they indicated something that would reject them, e.g. if after page 3 they indicated that they could not participate in the experiment without the use of glasses.

APPENDIX 4. METHOD

Figure 4.2: The figure shows the flow of the recruitment survey including what responses led to either qualifying or disqualifying for the experiment.



4.1.2 Experiment Design

Two documents were created for the experiment, an introduction to be read before the experiment started and the question sheets (QS), which were to be answered during the experiment.

Instruction document

The instruction document can be found in supplementary material 2.1. This document first had a section thanking for the participation and explaining that the participants had to build a LEGO set while filming themselves. It was also explained that the participants should make sure that they could set aside 2 hours without interruptions to complete the experiment. The document included a link and a QR code, both leading to an instruction video helping them prepare for the experiment. The participants were instructed to watch the video when they were ready to start the experiment. It was also explained what the video would show, including the structure of the experiment, being that they had to build 17 flowers in total and answer 5 questions after building each flower, which could be found in the QS. Lastly, the participants were told that they could contact any of the two authors of this study using the e-mail addresses in the document if they had any questions, and that they could keep the LEGO set as a reward for completing the experiment.

Instruction Video

The instruction video can be found in supplementary material 2.2. The purpose of making an instruction video was to make it easier for participants to set up for the experiment and to increase the likelihood that they would record and fill out the QS correctly. The video was intended to guide the participants through completing the experiment, and it started by showing the materials needed, which was the LEGO set, the instruction document, the QS, a pair of scissors, a pencil, and a recording device (e.g. a smartphone or a PC). The next part of the video instructed the participants to find a quiet room and to sit down at a table with enough space to fit the materials. Here, the participants were reminded to remove their glasses if they were wearing any and to secure their hair behind their ears or put it up in a ponytail, if they had long hair. The

next part of the video instructed the participants to put their phone on *Do not disturb*-mode and to make sure that there were at least 80% battery on their recording device.

After this, there was a pause in the video, where the participants were encouraged to set up as instructed before watching the rest of the video. Here, it was stated that they should not open the LEGO box yet. The next part of the video was meant to help the participants set up their device to record the experiment. Parvanta et al. (Parvanta et al., 2022) made a short video to demonstrate optimal body and face positioning to maximize data collection, since they conducted an experiment using a remote test setup. We decided to do the same for our experiment, which is why this part of the video showed how to position the camera, so their face, arms, hands and a part of the table would be visible before starting the recording. The participants were encouraged to lean a bit to both sides to make sure that their face would still be within the frame, e.g. when they had to fill out the QS. This part also showed how the lighting should be during the recording, illustrating that the light should not come from behind, since it would create shadows on the face, and that the room should not be too dark, since it would make it difficult to see their face. Lastly, it was shown that the camera should not be placed to the side, but that it should face the participants directly from the front. At the end of this part, the optimal camera positioning and lighting was shown.

After the setup part, the video guided the participants through the experiment by first telling them to read the instruction document, if they had not already done so. When they were ready to begin the experiment, they were instructed to find the QS and to fill out the first page after familiarizing themselves with the document. After this, the participants were instructed to open the box and take out the content. Here, they were also told that they should follow the manual from the box during the experiment. Before building, the participants were instructed to fill out the second page of the QS based on the unboxing experience. Then the participants were instructed to start building by opening the manual and build as they naturally would at their own speed. During the building process, the participants were instructed to build one flower at a time, and that some flowers had to be built more than once. Furthermore, the participants were instructed to not talk while building based on recommendations by Shah et al., 2013 who state that automatic detection of AUs is not possible while a person is talking, because it disturbs the software. The participants were instructed to show the flower to the camera

every time they finished building a flower, and then to fill out the corresponding page of the QS, where the name of the flower would be written at the top. The participants were of course allowed to take breaks during the experiment, but were instructed to finish building a flower before taking a break and to stop or pause the video during the break.

The next part of the video was about the debriefing, where the participants were informed that the last page of the QS had some final questions, that the participants were asked to repeat out loud and then answer in either Danish or English. After this, the participants were instructed to stop the recording, since the experiment would be over, and to send the materials back to us. Because the video turned out to include a lot of information, which could be difficult to remember, a summary section was added at the end of the video summarizing the most important information. These were that the participants should build one flower at a time and that they should respond to the corresponding questions in the QS after building each flower. The last piece of information in the video was that the participants could rewatch the video as many times as they wanted and to contact us if they had any questions before starting the experiment.

Question Sheet

The QS was the document that the participants had to fill out during the experiment. An example of the QS can be found in supplementary material 2.3. The first page of the QS was the preparation page, that the participants had to fill out before opening the LEGO set. Here, the participants had to make sure that they could answer yes to the following statements by checking boxes accordingly:

- Are you alone in the room?
- Is it quiet in the room (no music, other people talking etc.)?
- Have you removed your glasses (if you were wearing any)?
- Have you removed your hair from your face by putting it behind your ears or putting it up in a ponytail (if you have long hair)?
- Do you have at least 10 GB of space free on the device that you will use to record?
- Did you put your phone on 'do not disturb' or 'flight mode' (if you're using your phone to record)?
- Do you have at least 80% battery power on your phone (if you're using your phone to record)?

These statements were all mentioned in the introduction video as well. On the same page, the participants had to assess their current mood by filling out a 7-point scale regarding enjoyment, frustration, challenge, boredom, and excitement, where 1 was labeled *low*, 4 was labeled *neutral*, and 7 was labeled *high*. The 5 parameters were chosen based on Čertický et al. (2019), who measured enjoyment and Mandryk et al. (2006), who measured frustration, boredom, challenge and excitement, and retrospectively enjoyment. Mandryk et al. (2006) investigated the connection between subjective ratings and measurements of galvanic skin response, heart rate, electrical activity and respiration in the heart among others. FEA was, however, not measured. In their experiment, the participants were playing a hockey game either against the PC on different levels or against another player, while assessing 3 parameters after each condition. Čertický et al. (2019) measured heart rate, respiration and electroencephalography, among others. Once again, FEA was not measured. Here, different PC games were played, while the screen and the participants were recorded, which was used retrospectively to get the participants to rate their enjoyment throughout the game session. Mandryk et al. (2006) also measured fun, but we chose to not include this in the experiment as fun and enjoyable were assumed to be semantically connected in the chosen context of the experiment. Excitement could also be argued to be similar to enjoyment, but an assumption was made that excitement would be more focused on expectations of what is next to come, rather than being focused on the present enjoyment. These parameters were discussed with our company contact at LEGO, who agreed with these assumptions and thought they were interesting parameters to investigate in the given context.

The 7-point scale was chosen to be able to measure as much variation in the participants' subjective assessment as possible without making the scale too complicated and time consuming to fill out as the participants had to do so several times during the experiment. The participants had to fill out the scale for the 5 parameters 20 times in total; assessing their current mood before opening the box, assessing the unboxing experience before building, assessing the building experience of each of the 17 flowers in the LEGO set, and assessing the overall experience of the experiment after having built all 17 flowers. For all assessments except for the first one, a picture was shown on the page related to what the participants had to assess. The assessments based on each of the 17 flowers had an image of the flower as it would look when completed. An exam-

ple of this can be seen on Figure 4.3, which shows the page where the participants had to assess the 5 parameters based on *Welsh poppy 1*. The name of each of the different flower species could be found in the instruction manual from the LEGO set. The order of the flowers in the QS followed the order of the instruction manual from the LEGO set, which was meant to make it easier for the participants to find the assessment scales for each flower. As seen in Figure 4.3, some of the flowers were enumerated as well, which only applied for the flowers that the participants had to build more than once.

The assessment based on the overall experience of the experiment had an image of all 17 flowers as a bouquet along with the following description: *"Rate the following based on your **overall experience** of the experiment. Please consider both aspects of participating in the experiment (e.g. setup) as well as aspects of the building process"*. The last page of the QS was the debriefing page, where the participants were instructed to answer some final questions by saying the answer out loud. They were also told to elaborate as much as they could, and that they could answer in either English or Danish. The first two questions on the page were as follows:

- What did you think about participating in the experiment?
- What did you think about building the LEGO set?

The purpose of these questions were to gain insights that could be used to explain e.g. the participants' assessments throughout the experiment and their answers to the questions that followed. For the following questions, the participants were again asked to consider both the aspect of participating in the experiment as well as the building process. The next questions were as follows:

- Were there moments you enjoyed?
 - If yes, which moments did you enjoy the most?
- Were there moments where you felt frustrated?
 - If yes, which moments did you feel the most frustrated?
- Were there moments where you felt challenged?
 - If yes, which moments did you feel the most challenged?
- Were there moments where you felt bored?

Figure 4.3: The figure shows an example of a an assessment page where the participants were asked to rate the 5 parameters based on having built *Welsh poppy 1*.

Welsh poppy 1



Rate the following based on the steps you followed to build welsh poppy 1.

Enjoyment

Low

Neutral

High

--	--	--	--	--	--	--

Frustration

Low

Neutral

High

--	--	--	--	--	--	--

Challenge

Low

Neutral

High

--	--	--	--	--	--	--

Boredom

Low

Neutral

High

--	--	--	--	--	--	--

Excitement

Low

Neutral

High

--	--	--	--	--	--	--

- If yes, which moments did you feel the most bored?
- Were there moments where you felt excited?
 - If yes, which moments did you feel the most excited?

These questions were meant to be validation questions, meaning that they could be used to see if for instance the participants mentioned the flower that they had rated the highest for frustration, when asked about a moment where they felt most frustrated. This could also be used to see if participants often mentioned the first or last moments in the experiment, which could mean that the answers were affected by the primacy or recency effect (Nordfang and Nørby, 2017, p. 152). Lastly, the participants had the opportunity to give additional comments, before they were told that they could end their recording.

The order of the 5 parameters that the participant had to assess were balanced using a latin square. To make the assessments easier for the participants and to limit the number of mistakes made while assessing the parameters, the order of the parameters were the same throughout the QS for each participant, meaning that the order was only balanced between the participants. This was done by first making a standard 5x5 Latin square, where a, b, c, d, and e are inserted as placeholders in the first row and in the first column. In this matrix, the placement of the placeholders are the same for rows and columns with the same number, e.g. second row and second column, but the placement of the placeholders are moved by one each time they are placed in another row or column. This means that a, b, c, d, e in the first row and first column, turns into b, c, d, e, a, in the second row and second column and so on. After making the standard matrix, the columns were randomized, by inserting the numbers from 1 to 5 in a random generator. After changing the order of the columns, the same procedure was done for the rows. This was then the 5 orders that were chosen, which can be seen in Table 4.1. We chose to use only these 5 orders, which means that the number of participants had to be a multiplication of 5.

4.1.3 Pilot Study

A pilot study was conducted with a 48 year old woman as participant. She was instructed to read the instruction document, watch the instruction video and carry out the experiment, but only by opening the bags with number 2 on them in the box, as the bags

APPENDIX 4. METHOD

Table 4.1: The table shows the Latin square used for the order of parameters in the QS. *Order* indicates the order number and *Row/Column* indicates the column number and row number from the standard Latin square.

Order	Row/Column	3	5	1	2	4
1	5	Enjoyment	Frustration	Challenging	Boredom	Excitement
2	1	Excitement	Challenging	Boredom	Enjoyment	Frustration
3	4	Boredom	Excitement	Frustration	Challenging	Enjoyment
4	3	Challenging	Enjoyment	Excitement	Frustration	Boredom
5	2	Frustration	Boredom	Enjoyment	Excitement	Challenging

with number 1 had been used to record the instruction video and another copy of LEGO set 10313 was not available at the time. As the purpose of this pilot study was to test the instructions, we decided to only let the participant build the flowers from the bags with number 2 on them to shorten the experiment. The debriefing was also included in the pilot study.

The pilot study resulted in some changes to the instructions. First of all, it led to including the summary part of the instruction video, as the participant commented that there was a lot of information which made it difficult to remember and to know what was most important. Moreover, the pictures were added to the QS as they were not included at the time of the pilot study, which resulted in the participant not answering the questions for the correct flowers after building a flower. Here, it is important to note that the participant had received a QS with all flowers included even though the participant only needed to build the flowers from the bags with number 2 on them. This meant that she had to go only by the name of the flowers which was difficult for the participant as she also had to ignore the first pages with questions for the flowers from the bags with number 1 on them. The participant from the pilot study was not accustomed to build with LEGO and had not done so for years, which means that the set was difficult to build and that also made the experiment difficult to complete for her.

These challenges resulted in the pilot study taking 40 minutes even though only approximately 1/4 of the set was built. Furthermore, the participant did not build all the flowers from the bags with number 2 on them, because she did not realize that some of the flowers had to be build more than once. This resulted in adding the summary part in the video and the pictures to the QS, to make the experiment easier to complete. After having made the changes following the first pilot study and having received 20 copies

of set 10313, another pilot study was carried out. This time, the participant received a factory sealed box and was asked to complete the entire experiment. As the participant had no issue completing the experiment, this was not considered as a pilot study, but instead the individual was considered the first participant in the experiment.

4.1.4 Participants

20 people (9 women and 11 men) participated in the experiment. Table 4.2 shows demographic information such as gender, age, highest level of education completed and occupation. The table also shows how frequent they built with LEGO, if they build alone or with others, and their enjoyment towards LEGO on a scale from 1 (I don't like building with LEGO at all) to 7 (LEGO building is the most enjoyable thing I can think of). P03 and P14 had issues with recording, which led to only receiving usable data from 18 participants.

Table 4.2: The table shows demographic information about the participants from the experiment, as well as information regarding their relation towards building with LEGO. No data was received from P03 and P14. - indicates that no response has been collected because the ended due to prior answers.

ID	Gender	Age	Education completed	Occupation	Building frequency	I build with LEGO...	LEGO Enjoyment
P01	Male	26-30	Intermediate higher education	Working full time	It has been years	For the sake of others	5
P02	Female	26-30	Long higher education	Working full time	Never	-	-
P03	Female	26-30	Intermediate higher education	Working full time	Once a year	As a social activity	6
P04	Male	26-30	Intermediate higher education	Working full time	Once a year	As a social activity	7
P05	Female	18-25	Intermediate higher education	Gab year	Once a year	For the sake of others	4
P06	Male	26-30	Intermediate higher education	Working full time	Once a month	Alone	7
P07	Female	26-30	Long higher education	Working full time	It has been years	As a social activity	5
P08	Male	26-30	Long higher education	Working full time	It has been years	I don't build with LEGO	-
P09	Female	26-30	Intermediate higher education	Student/trainee	Once a year	For the sake of others	5
P10	Female	18-25	Intermediate higher education	Student/trainee	It has been years	I don't build with LEGO	-
P11	Male	18-25	Primary/elementary school	Working full time	It has been years	As a social activity	5
P12	Female	18-25	Intermediate higher education	Student/trainee	It has been years	I don't build with LEGO	-
P13	Female	18-25	High school (gymnasium)	Student/trainee	It has been years	Alone	5
P14	Male	18-25	High school (gymnasium)	Student/trainee	It has been years	Alone	5
P15	Male	26-30	Short higher education	Student/trainee	It has been years	Alone	5
P16	Male	26-30	Intermediate higher education	Student/trainee	Once a year	For the sake of others	5
P17	Male	18-25	Intermediate higher education	Student/trainee	Once a year	Alone	6
P18	Male	26-30	Intermediate higher education	Working full time	It has been years	Alone	6
P19	Female	18-25	High school (gymnasium)	Working part time	It has been years	Alone	7
P20	Male	26-30	Long higher education	Seeking a job	Once a year	Alone	5

4.1.5 Procedure

The instruction document, the LEGO set, and the QS with the designated order for the 5 parameters were delivered to the participants. They were instructed to complete the experiment and return the recorded video as well as the QS responses within 4 weeks, meaning that they would have enough time to be able to find a fitting time to set aside for building. The participants first read the instruction document and watched the instruction video. Afterwards, they set up their device and started recording. As the next step, they filled out the first page of the QS, then opened the LEGO box and filled out the second page of the QS. Next, the participants followed the LEGO manual by building one flower at a time after which they answered the QS for the corresponding flower. This step was repeated for all 17 flowers in the LEGO set. After having built the entire set, the participants filled out the page in the QS about their overall experience of participating in the experiment. Finally, they answered the debriefing questions and stopped the recording. The video of the participants completing the experiment and their assessments from the QS was then sent to us to be analyzed. In total, the experiment lasted between 90-180 minutes for each participant.

5 Results and Analysis

From the experiment, videos lasting between 90-180 minutes and the assessments of the 5 parameters for each step were received from 18 of the participants. Data collection lasted about 4 weeks, as participants returned the video and responses to their question sheets (QS) over this time. Continuously, we prepared and analyzed the videos through OpenFace whenever we received the materials. The purpose of this chapter is to investigate whether the collected data is applicable to train models based on the action units (AU) from the OpenFace analysis to predict the subjective assessments of the five parameters. The process of analyzing the videos through the OpenFace software will also be explained.

5.1 Data Preparation

The preparation of data involved watching and familiarizing ourselves with how the participants completed the steps by watching the videos. Participants generally followed the instructions from the video and the LEGO manual quite well. Only few building errors were made, and the only deviations from the experiment instructions were that three participants once forgot to fill out the QS between steps, but in those cases, they would always remember after the subsequent step, explain verbally, and fill out the QS based on their actual experience with the step. Two participants also forgot to record the unboxing step and a single participant had issues with recording the step with the last flower.

As part of watching the videos, we also noted down timestamps of the beginning and end of each step. By doing this, we could edit the video to only include the unboxing and the building steps before running the OpenFace analysis, so that no irrelevant

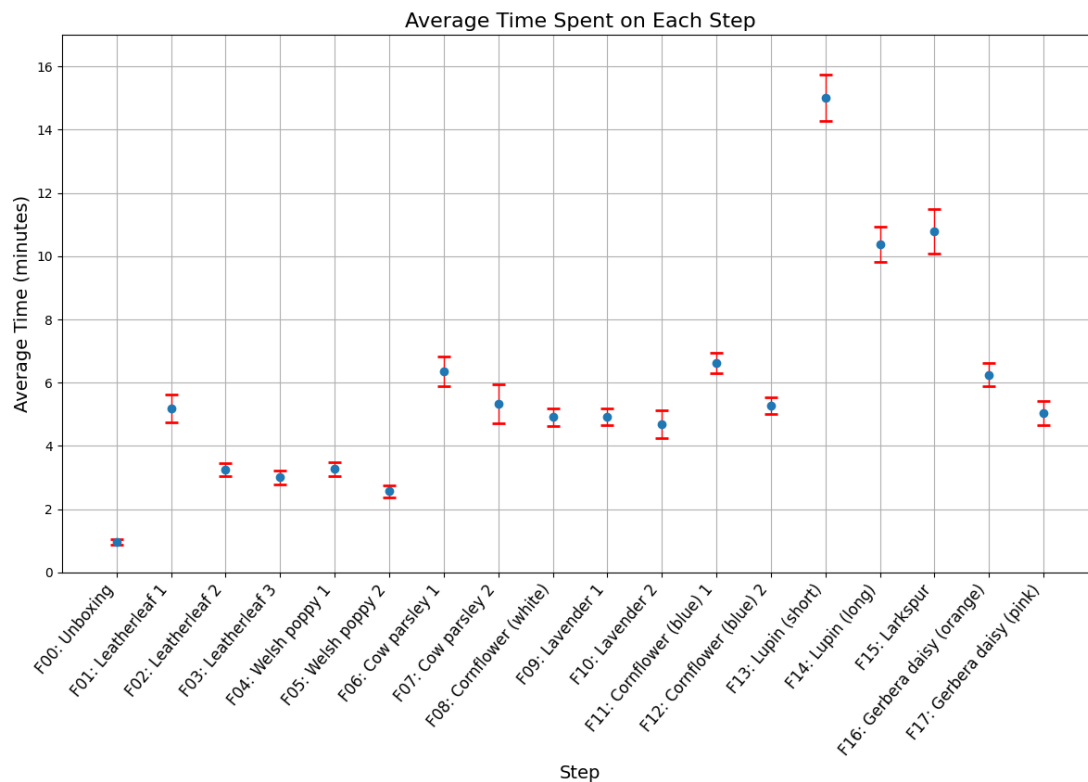
parts were analyzed. This was mainly done prior to analysis because running each video through the software took about 4 to 5 times as long as the length of the video. Specifically, participants filling out the QS, breaks, and participants unpacking plastic bags in between steps were edited out of the video. The latter was left out of the analysis videos as this was done 4 times during the experiment, and the bags included LEGO bricks for 2 to 7 flowers each, and as the participants were asked to rate their experience of building the specific flower, we assumed the unpacking of bags would not be part of their rating. Either way, this step typically lasted 30 seconds or less and is not assumed to have any major impact on the later conducted analysis.

A majority of participants had a tendency to smile when they showed the flower to the camera after finishing each step, possibly because of being proud and happy with finishing their creation. Another possibility is, however, that the smile was stemming from seeing themselves in the camera and/or knowing that we would be watching the videos later on. Due to the potential of the latter and this not being a genuine smile related to the building experience, we decided to edit this out as well. Finally, the videos were set to 300% their original speed. This was due to OpenFace already sampling at 30Hz, and this way we would effectively reach ≥ 10 frames per second, but still reaching a substantially faster run time in OpenFace, which was necessary due to the many hours of video material. With ≥ 10 frames per second, we assumed that micro-expressions would still be captured, since these can last up to 0.5 seconds (Paul Ekman Group, 2024). The video preparation resulted in 39 hours and 42 minutes of video material across the 18 participants, amounting to 13 hours 14 minutes in 300% speed.

5.2 Time Spent on Steps

Figure 5.1 shows an overview of the average time, participants spent on each step. The least time was spent unboxing, followed by some of the first steps, whereas F13 followed by F14 and F15 took the longest to build.

Figure 5.1: The figure shows mean time spent on each step with error bars showing standard error.

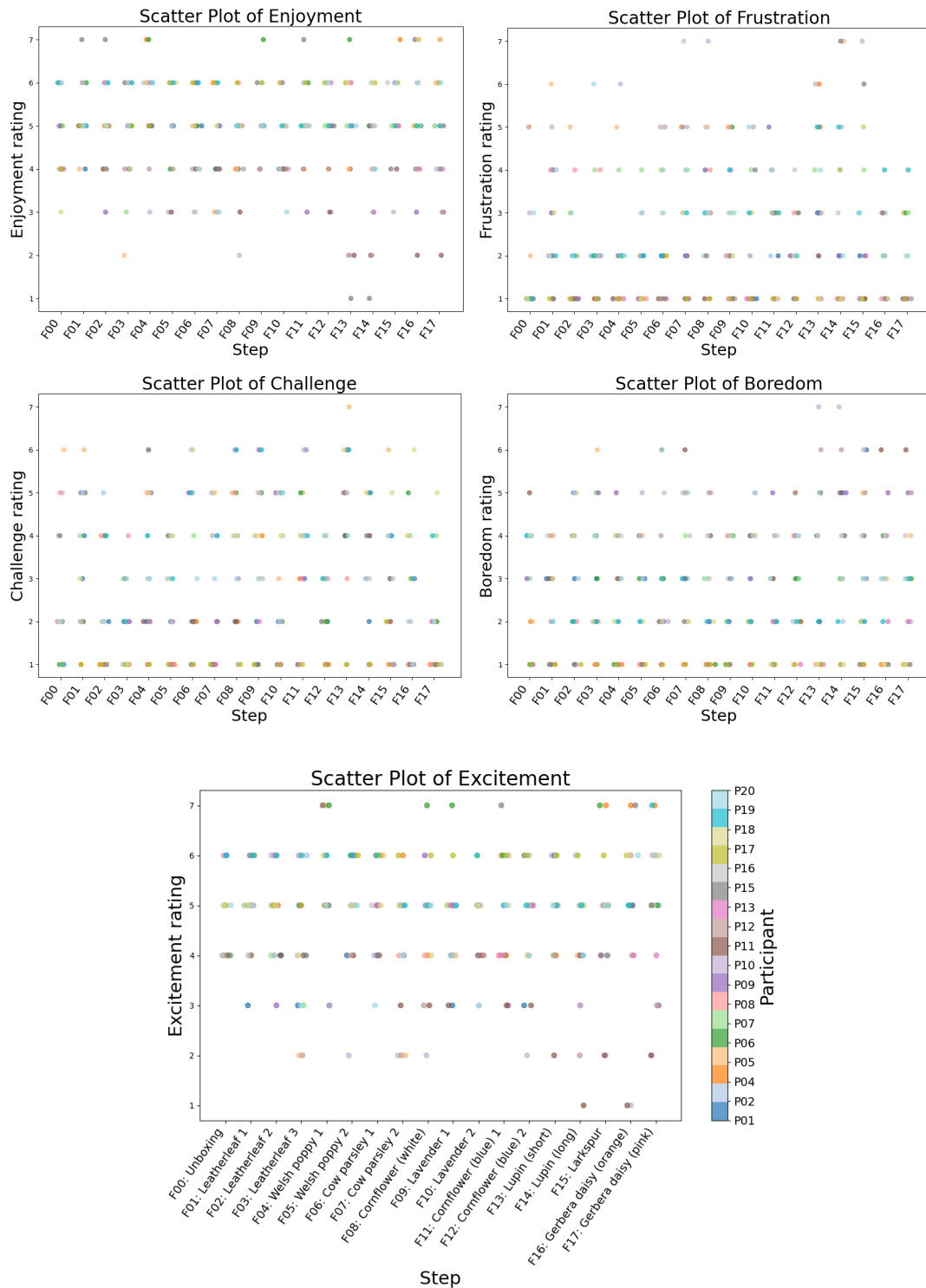


5.3 Participant Ratings in Question Sheets

To get familiar with the participants' ratings in the QS, the data can be seen plotted in Figure 5.2. Here, it can be seen that the participants generally used higher ratings of 4-6 for *enjoyment* and *excitement* as well as lower ratings from 1-4 for *frustration*, *challenge*, and *boredom*.

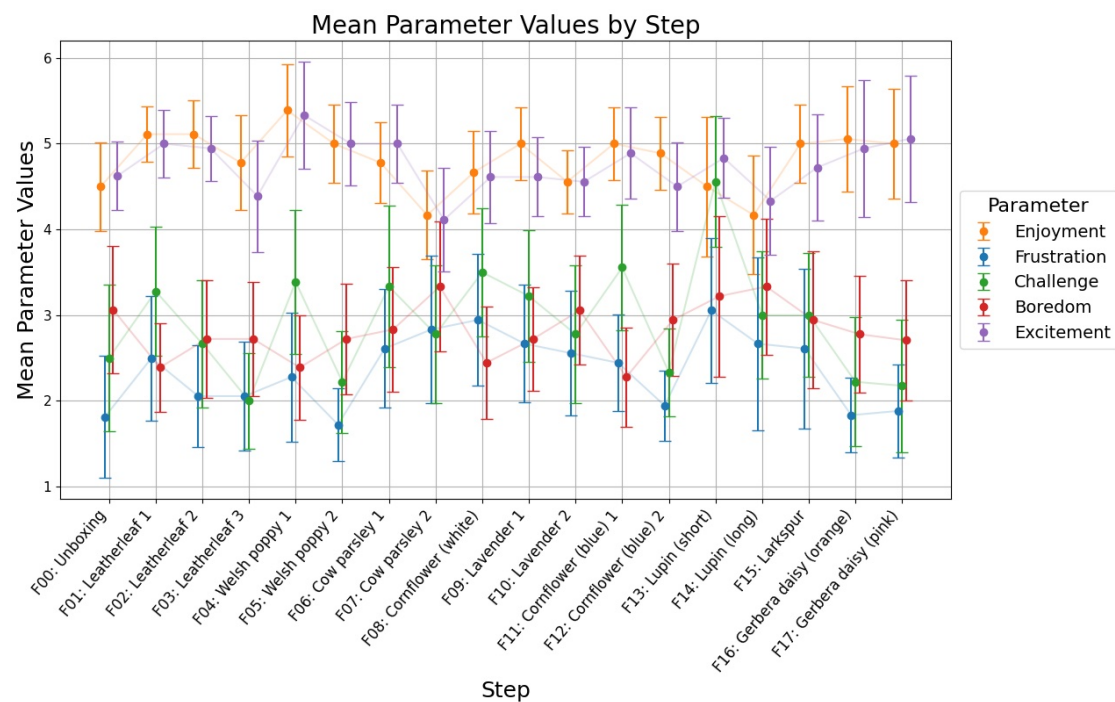
APPENDIX 5. RESULTS AND ANALYSIS

Figure 5.2: The figure shows scatter plots of how each participant have rated each of the parameters for every step.



Looking at Figure 5.3, we see the mean assessment values for each parameter in each step given by all participants. Here, it confirms what is seen in Figure 5.2 that *enjoyment* and *excitement* were ranked quite similarly, and also the highest among all 5 parameters with mean values between 4 and 6. *frustration*, *challenge*, and *boredom* were ranked a bit lower with mean values between 1 and 4.

Figure 5.3: The figure shows the mean values for each parameter for each step. Error bars show 95% CI.



5.4 OpenFace Analysis

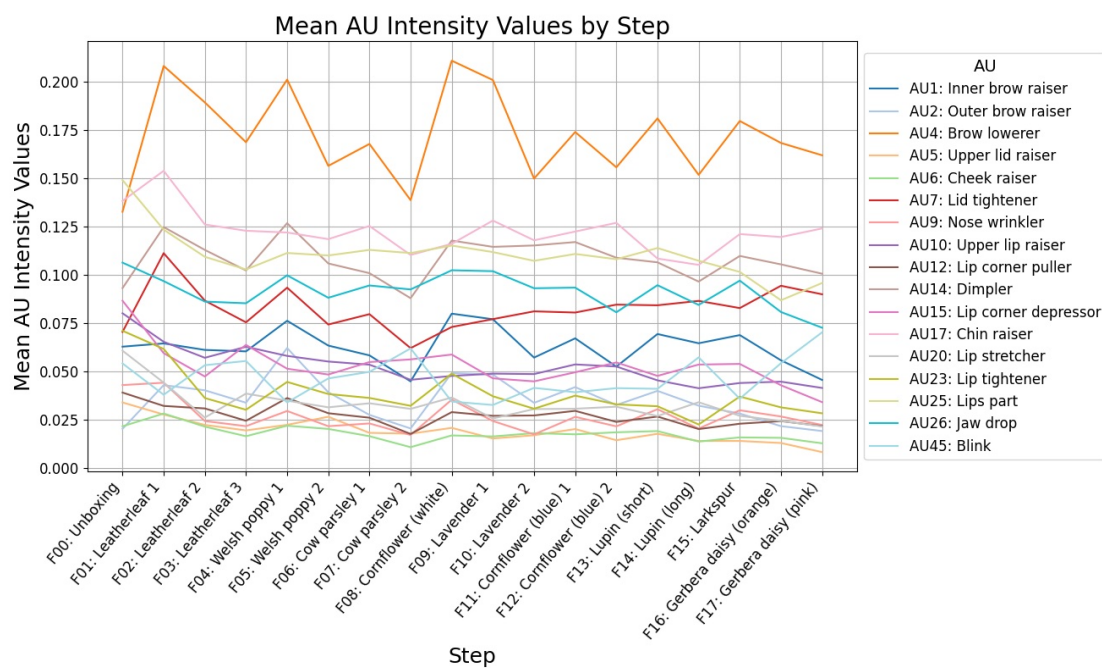
Following the preparation of videos, we analyzed each video using the OpenFace software explained in appendix 3.2. After following the installation guide and installing the toolkit, the `FeatureExtraction [video path]` command was run for each video (Baltrusaitis, 2021; Baltrusaitis, 2019). The output of this included HOG files, a video identifying the facial features as well as a .csv file containing the x and y coordinates for the facial features from the detected landmark as well as values for *presence* and *intensity* of AUs for each frame. Each row in the .csv file corresponds to one frame in

the video, and is referred to as a data point. For each data point, the output also included a *success* column indicating whether a face was detected as well as a *confidence* row indicating from 0-1 how certain the prediction of facial features and AUs are.

5.4.1 Post-processing of OpenFace Output

Post-processing was done for every output file of the OpenFace analysis. This initially involved excluding all data points where confidence fell below 0.85, following the approach of Krause et al. (2020), who also analyzed videos using OpenFace 2.0, as well as those with success being 0, indicating that either no face was detected or that it was not properly detected. This resulted in removing 2.20% (varying from 0.21% to 8.08%) of the total data points, leaving us with 1,035,918 usable data points with *presence* and *intensity* values for each AU across all 18 participants in all 18 steps. The data based on AU *intensity* values was distributed as indicated in Figure 5.4.

Figure 5.4: The figure shows the mean *intensity* normalized to range from 0-1 for the 17 AUs for each step.



Next, data was processed in different ways before feeding data to train the models. Several different approaches to formatting the data were prepared to attempt different

combinations in order to fit the most useful models. Table 5.1 explains the different approaches.

Table 5.1: The table shows the different approaches to formatting data that were used to train the machine learning models. The formatting was done by processing AU *intensity* (I_{AU}) and *presence* (P_{AU}) values from the OpenFace output.

AU type	Term	Range	Mathematical Explanation	Description
Intensity	Normalized by max AU	0-1	$I_{normMax} = \frac{I_{AU}}{5}$	Intensity of AUs normalized by dividing with 5, the max value any AU can reach.
Intensity	Normalized by individual participant	0-1	$I_{normInd} = \frac{I_{AU}}{I_{indMax}}$	Intensity of AUs normalized by dividing with the max value of the given AU for the given participant.
Intensity	Mean overall	0-1	$I_{meanOverall} = \frac{1}{N} \sum_{i=1}^N I_{AU_i}$	Intensity mean value per step per participant for “normalized by max AU” values.
Intensity	Mean if present	0-1	$I_{meanPresent} = \frac{1}{N_p} \sum_{i=1}^{N_p} I_{AU_i}$	Intensity mean value per step per participant for “normalized by max AU” values excluding 0 values, so only including intensity values when the AU is present.
Presence	Mean	0-1	$P_{mean} = \frac{1}{N} \sum_{i=1}^N P_{AU_i}$	How often AU is present per step per participant.
Presence	Significant from binomial distribution	0, 0.5, 1	$P_{sig} = \begin{cases} 1 & 0.5 < CI_{95\%} \\ 0.5 & 0.5 \in CI_{95\%} \\ 0 & 0.5 > CI_{95\%} \end{cases}$	If presence mean is significantly more or less present than not per step per participant with 95% CI.
Presence	Dynamics	0-1	$P_{dynamics} = \frac{\text{Number of changes}}{N-1}$	How often AU changes between being present and not present.

Since *intensity* of AUs was ranked from 0-5 in the output, we normalized this to range from 0-1 following recommendations from Singh and Singh (2019), so all variables would be within this range. This was done both in relation to the maximum value any AU could reach, being 5, and to the individual maximum value the specific AU reached for the given participant. The latter was done since some participants never reached the maximum value within the entire experiment for some of the AUs. This could either mean that these participants were not affected enough during the experiment to reach the maximum *intensity* of these AUs or that they are simply less expressive and therefore would never reach it. Either way, it had the potential of skewing the data to make some participants whose AU values were more explicit have too big of an impact on the training data compared to others with less explicit AU values. For this reason, both the *intensity* values were normalized using both approaches. Furthermore, mean

values with standard deviations (SD) were calculated for each step for each participants. This was done for all values (including 0) as well as for only values above 0, indicating the mean *intensity* of the AU only while being present.

Mean and SD was also calculated for *presence* values per participant per step, indicating the proportion of time, the given AU was present. For this, 95% confidence intervals (CI) were determined, to then assign either 1 or 0, depending on whether the AU was significantly more present than not, or the other way around. If 0.5 was part of the CI, the data point was assigned 0.5.

5.5 Training Models

The first step of training the models involved including the mean values of *intensity* and *presence* of AUs by different approaches. 5 models were intended to be trained, with the target variables being one of the 5 self-assessment parameters in each. Table 5.2 shows the results for 5 models trained using the data for *intensity mean overall* (based on *normalized by max AU data*) and the corresponding SD, as well as *presence significant from binomial distribution* and *dynamics* as feature variables.

Random forest regression as well as random forest classification was run in Python (V3.12.2) with the commands `RandomForestRegressor(random_state=42)` and `RandomForestClassifier(random_state=42)` from the *sci-kit-learn* package (Scikit-learn, 2024a; Scikit-learn, 2024b). `random_state` was set to 42 for reproducibility. The scripts with corresponding data files for all attempts described in this section can be found in supplementary material 3.1.

Before training the models, we researched which methods were applicable. Both regression and classification seemed like appropriate first methods to try out, depending on whether we consider the self-assessment parameters to be on an ordinal or interval scale. The topic of how to treat self-assessed scale data has been discussed widely, but research has found that random forest regression models show similar results for interval and ordinal type data (Janitza et al., 2016). Random forest is a typical method to initially attempt for training models, as it provides relatively high interpretability compared to other methods such as neural networks as it provides feature importance (Breiman, 2001). Further, it is still able to handle complex interactions between variables as op-

APPENDIX 5. RESULTS AND ANALYSIS

Table 5.2: The table shows the evaluation metrics for the 10 models run, 2 for each 5 self-assessment parameter, for Random Forest regression and classification respectively. The feature variables are based on **means** of *intensity* and *presence* AU values. MAE refers to Mean absolute error, MSE to mean squared error, and RMSE to root mean squared error.

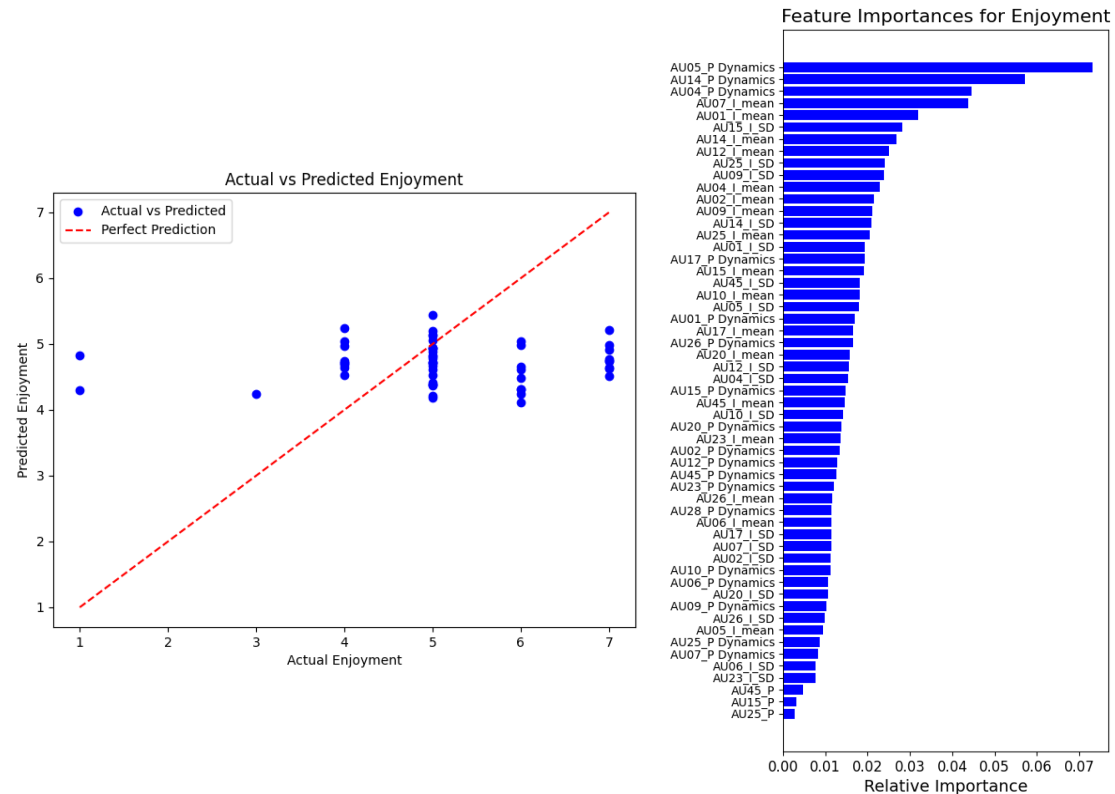
Target variable	Regression Random Forest				Classification Random Forest			
	MAE	MSE	RMSE	R ²	Accuracy	Precision	Recall	F1 Score
Enjoyment	1.05	1.88	1.37	-0.14	0.31	0.14	0.16	0.15
Frustration	1.3	2.33	1.53	-0.09	0.31	0.08	0.13	0.10
Challenge	1.42	2.68	1.64	0.06	0.24	0.17	0.19	0.15
Boredom	1.57	3.23	1.8	-2.78	0.13	0.15	0.11	0.09
Excitement	0.88	1.29	1.13	-0.43	0.31	0.20	0.23	0.21

posed to e.g. linear models, and the fact that it uses a combination of multiple decision trees, makes it resistant to overfitting (Cutler et al., 2007; Biau, 2012). It is, however, computationally heavy, especially with large data sets (Probst et al., 2019; Bergstra and Bengio, 2012). Finally, Random Forest classification allows for classifying into multiple categories, where some other methods only allow for two categories, which would have led us to lose quality in the data by reducing the categories into e.g. positive and negative for each self-assessment parameter (Hastie et al., 2009).

A common way of splitting data into training and test data sets is a random split with 70% for training and 30% for test (Xu and Goodacre, 2018). Because of dependency in our data since many data points come from the same participant, we instead decided to exclude a set of random participants from the training set. Initially, we used P6, P15 and P18, but also tried a couple of other variations, with no substantial difference in results. Different approaches of mean were attempted as well, including using *normalized by individual participant data* as well *intensity mean if present data* (see table 5.1). Neither of these attempts yielded substantially better results.

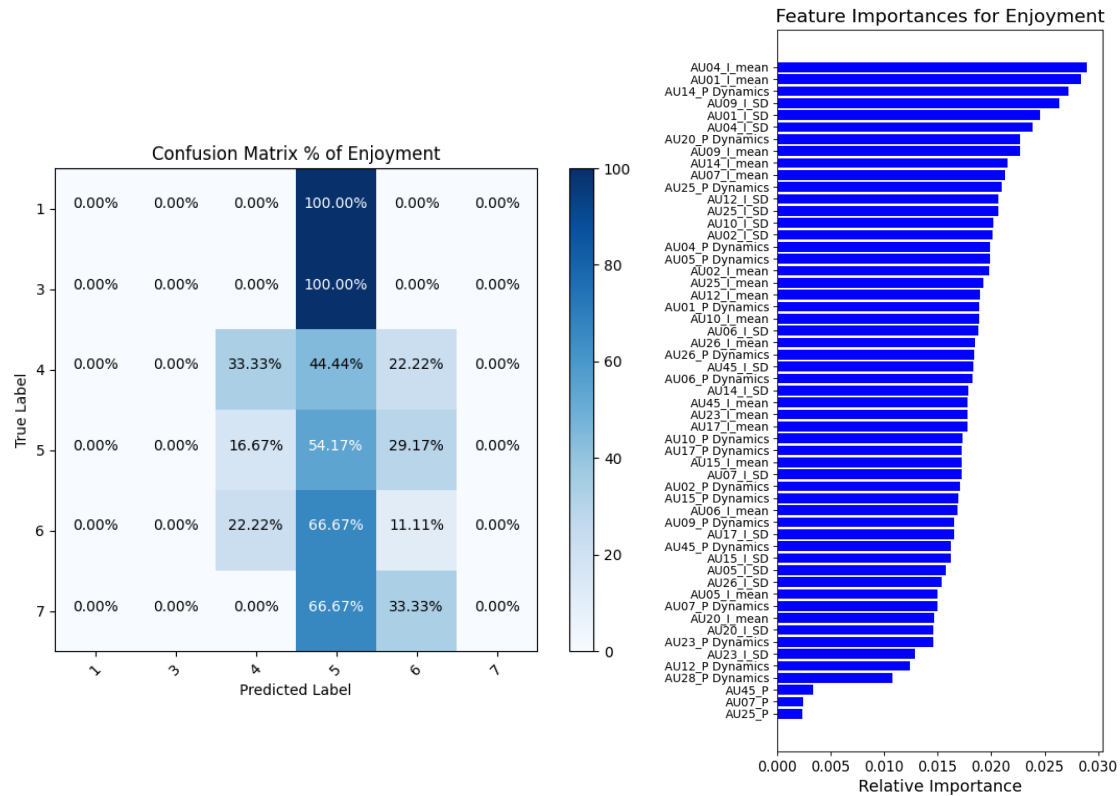
An example of a plot for regression data and the confusion matrix for the parameter *enjoyment* can be seen on Figure 5.5 and 5.6. These represent the results shown in Table 5.2. The figures also show the importance of each feature in the models trained for *Enjoyment*. Based on this, we also attempted to exclude a number of the least important features, however, once again without any substantial change in the results. The regression plots, confusion matrices, and feature importance plots for the remaining 4 parameters can be seen in supplementary material 3.2.

Figure 5.5: On the left, the figure shows a visualization of the Random Forest **regression** model for *enjoyment* using **mean** values of *intensity* and *presence* AU values. To the right, the importance of the 55 most important features is indicated.



Since each step per participant usually included 1,000+ data points, reducing all these data points to one or two data points, e.g. mean and SD, meant that the training data frame only included 227 data points and the test data frame included 54 data points. This could be considered a quite small data frame for machine learning, since a common rule in machine learning is to have 10 times, or preferably 100 times, as many data points as the number of features (Maxwell et al., 2018). With only 227 data points in the training set, we would not be able to include all available features, and even though we attempted to only include 10, 15, and 20 of the most important features, the model performance did not improve. Therefore, since the collected data set was a lot larger, and the applied approach potentially compromised patterns in the data, the next step involved attempting to utilize all the collected data points directly in the model.

Figure 5.6: On the left, the figure shows a visualization of the Random Forest **classification** model for *enjoyment* using **mean** values of *intensity* and *presence* AU values. To the right, the importance of the 55 most important features is indicated.



5.5.1 Training Models Utilizing Individual Data Points

This approach included applying the self-assessment parameter values directly to each data point with the corresponding participant and step number. This increased the data frame to include approximately 901,935 data points in the training set and 157,252 data points in the test set. These numbers did, however, vary slightly depending on which participants were included in the test set, as the participants who spent longer on the experiment had more data points.

Regression and classification were again run with the RandomForestRegressor and RandomForestClassifier. However, they were both run through the command: RandomizedSearchCV(estimator=ranfor, param_distributions=param_dist, n_iter=20, cv=3, n_jobs=-1, verbose=2, random_state=42) with rf referring

to the regressor and classifier commands, and with `param_dist` defining the hyperparameters as:

```
param_dist = {
    'n_estimators': randint(50, 200),
    'max_depth': [10, 20, None],
    'min_samples_split': randint(2, 10),
    'min_samples_leaf': randint(1, 5),
    'max_features': ['auto', 'sqrt']
}
```

The purpose of using `RandomizedSearchCV` was to find the best combination of hyperparameters for the model (Scikit-learn, 2024c). As previously mentioned, Random Forest is a computationally heavy method, so instead of exhaustively searching all possible combinations of hyperparameters, `RandomizedSearchCV` randomly samples a fixed number of hyperparameter combinations, which is computationally more efficient and can still yield good results. The intervals for each hyperparameter are based on current best practice.

The results can be seen in Table 5.3 using *intensity* data *normalized by individual participant* (see table 5.1 for description) with P6, P15, and P18 excluded from the training set. Once again, several approaches were tried out, using *intensity* data *normalized by max AU* as well as *normalized per individual participant*, and with different participants excluded from the training set. None of the attempts yielded substantially better results.

Table 5.3: The table shows the evaluation metrics for the 10 models run, 2 for each 5 self-assessment parameter, for Random Forest regression and classification respectively. The feature variables are based on **individual data points** of *intensity* and *presence* AU values. MAE refers to Mean absolute error, MSE to mean squared error, and RMSE to root mean squared error.

Target	Regression Random Forest				Classification Random Forest			
	MAE	MSE	RMSE	R ²	Accuracy	Precision	Recall	F1 Score
Enjoyment	1.21	2.62	1.62	-0.12	0.36	0.12	0.14	0.11
Frustration	1.43	2.97	1.72	-0.01	0.27	0.13	0.12	0.09
Challenge	1.56	3.10	1.76	-0.12	0.19	0.17	0.17	0.15
Boredom	1.57	3.22	1.79	-2.89	0.17	0.11	0.11	0.09
Excitement	0.87	1.22	1.10	-0.50	0.44	0.15	0.19	0.15

An example of a plot for regression data and the confusion matrix using individual data points with the corresponding feature importance plots for the parameter *enjoyment* can be seen on Figure 5.7 and 5.8. The regression plots, confusion matrices, and feature importance plots for the remaining 4 parameters can be seen in supplementary material 3.3.

Figure 5.7: On the left, the figure shows a visualization of the Random Forest **regression** model for *enjoyment* using **individual** values of *intensity* and *presence* AU values. To the right, the importance of the each feature is indicated.

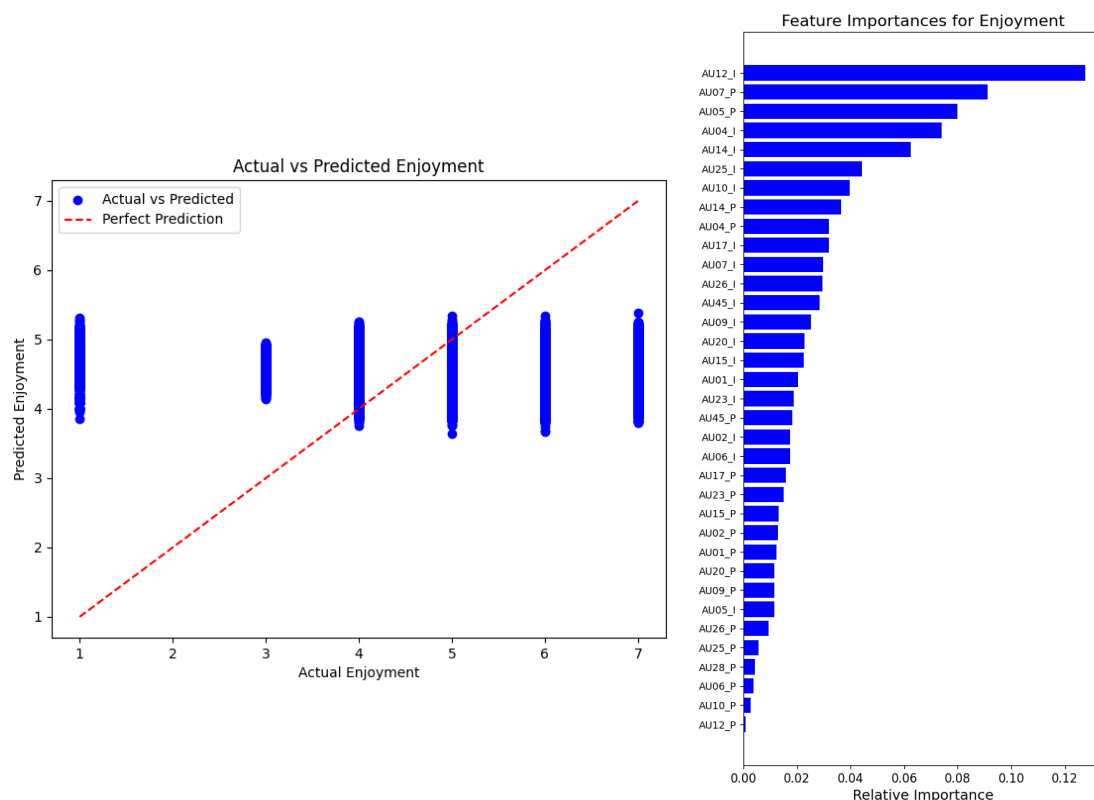
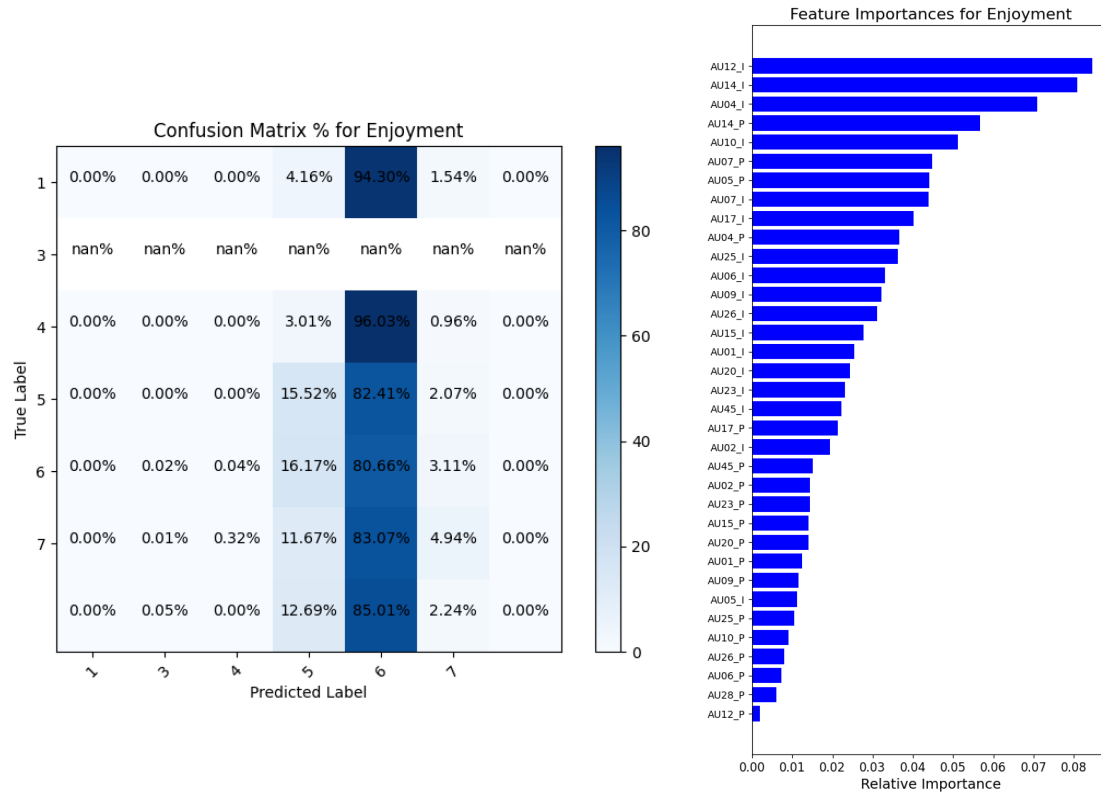


Figure 5.8: On the left, the figure shows a visualization of the Random Forest **classification** model for *enjoyment* using **individual** values of *intensity* and *presence* AU values. To the right, the importance of the each feature is indicated.



5.6 Intensity of AUs Compared with Self-reports

To get an overview of the connection between the activation of AUs and the ranking of the parameters from the experiment, this was plotted on Figure 5.9 and 5.10. Here it can be seen what AUs were active when the participants gave an assessment from 1-7 for the parameters *enjoyment*, *frustration*, *boredom* and *excitement*. On each of the figures, the AUs highlighted by thicker lines are assumed to have a higher intensity when the participants reported an increase in each of the four parameters (see Appendix 3.3 for more on these assumptions). *challenge* was not included in any figure as no literature was found regarding which AUs should be active when the participants experienced challenge.

APPENDIX 5. RESULTS AND ANALYSIS

Figure 5.9: The figure shows the activation of AUs dependent on the participants' *enjoyment* and *excitement* ratings. The bold lines indicate the AUs that were assumed to increase with increased ratings.

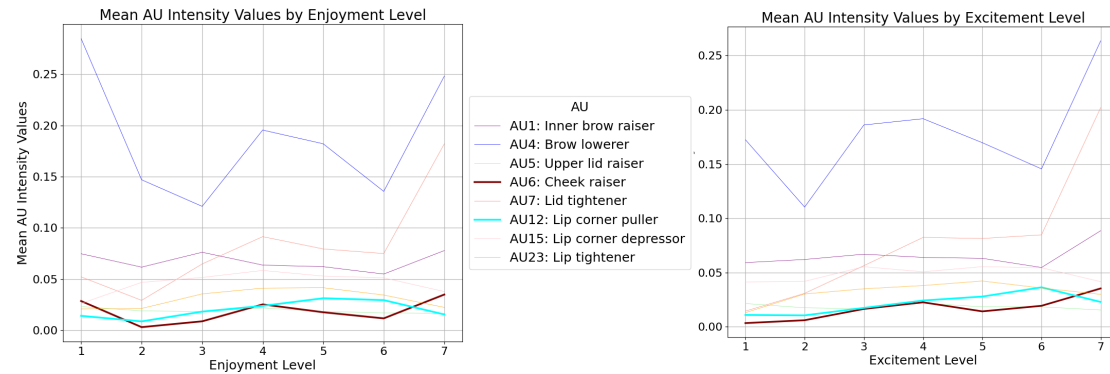
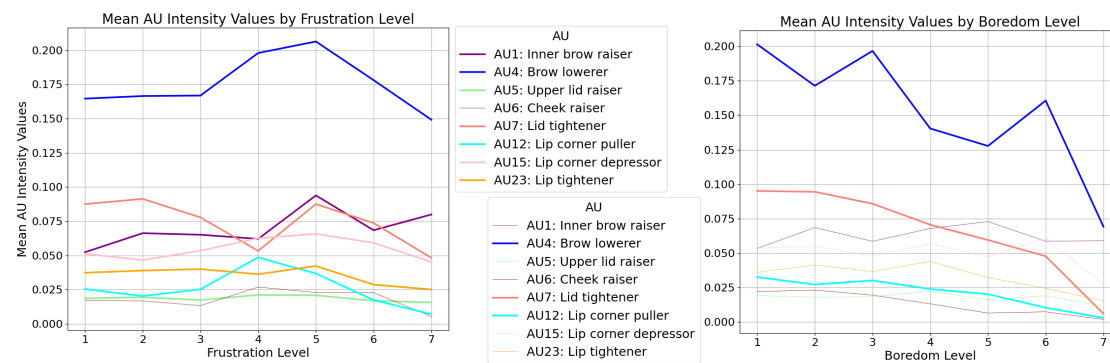


Figure 5.10: The figure shows the activation of AUs dependent on the participants' *frustration* and *boredom* ratings. The bold lines indicate the AUs that were assumed to increase with increased ratings.



6 Validation

This appendix describes an approach to validating the self-assessed parameters, in order to ensure that these were given consistently and depicted their actual experiences as closely as possible. Several issues have been described in literature with participant self-reports, and thus, this was an attempt to locate if and where these issues arose (Santoso et al., 2016; Čertický et al., 2019; Bañuelos-Lozoya et al., 2021). The validation consisted of two parts; looking at the participants' mean assessment for each parameter compared to their own overall assessment of the experiment, and looking at the similarity between the steps that the participants gave the highest assessments compared to the steps that the same participants mentioned in the debriefing part.

6.1 Mean and Overall Assessment Comparison

The last assessment that the participants were asked to give in the question sheet (QS) was an overall assessment of the experiment regarding the 5 self-assessment parameters. In the following, the participants' assessment of the 5 parameters for all steps will be compared to their overall assessments. Table 6.1 and Table 6.2 show the participants' mean assessment for each parameter and the corresponding overall assessment.

From the two tables, we see that in 41 out of 90 cases, the difference between a participant's overall assessment and their mean assessment through the experiment was > 0.50 , which is indicated by the green numbers. This includes 5 cases where the difference is 0.00, which in some cases was because the participants did not vary in their assessment of this specific parameter at all throughout the experiment. We chose to consider a category of values > 0.50 because the participants were not allowed to give assessments using decimal numbers, meaning that 0.50 also might as well indicate

APPENDIX 6. VALIDATION

Table 6.1: The table shows the general mood assessments indicated by *Initial [parameter]* and the mean assessment including standard deviation given by participant **P01 to P10** (except for P03) for each of the 5 parameters; *enjoyment, frustration, challenge, boredom, and excitement*. The table also shows the overall assessment that the participants gave for each of the parameters, which is compared to the mean for the same parameter resulting in a row indicating this difference. Differences of > 0.5 is indicated with green and differences of ≥ 1.5 is indicated with red.

	P01	P02	P04	P05	P06	P07	P08	P09	P10
Initial enjoyment	3	3	7	4	4	4	6	4	4
Mean enjoyment (SD)	4.21 (0.69)	5.21 (0.83)	5.26 (1.29)	4.68 (1.03)	5.16 (0.99)	4.84 (0.59)	4.89 (0.79)	4.63 (1.09)	3.68 (1.03)
Overall enjoyment	4	6	7	5	6	6	5	5	3
Difference	-0.21	+0.79	+1.74	+0.32	+0.84	+1.16	+0.11	+0.37	-0.68
Initial frustration	4	4	1	4	1	4	4	1	2
Mean frustration (SD)	2.32 (0.86)	2.42 (1.04)	1.68 (1.30)	3.79 (1.54)	2.00 (1.30)	3.68 (0.57)	3.16 (1.63)	2.16 (1.18)	3.63 (2.00)
Overall frustration	3	2	1	5	3	3	3	2	5
Difference	+0.68	-0.42	-0.68	+1.21	+1.00	-0.68	-0.16	-0.16	+1.37
Initial challenge	2	4	1	4	2	4	5	1	1
Mean challenge (SD)	2.16 (0.67)	3.26 (1.77)	1.84 (1.50)	4.00 (1.59)	2.42 (1.53)	4.00 (0.32)	3.58 (1.57)	3.74 (1.77)	3.74 (1.83)
Overall challenge	2	2	2	5	4	4	3	5	5
Difference	-0.16	-1.74	+0.16	+1.00	+1.58	0.00	-0.58	+1.26	+1.26
Initial boredom	4	5	1	4	1	4	1	3	4
Mean boredom (SD)	3.53 (1.04)	2.58 (0.88)	1.37 (0.58)	3.95 (0.94)	2.11 (0.97)	3.53 (0.60)	1.95 (0.10)	3.58 (1.18)	3.74 (1.74)
Overall boredom	4	3	1	4	3	3	2	3	5
Difference	+0.47	+0.42	-0.37	+0.05	+0.89	-0.53	-0.05	-0.58	+1.26
Initial excitement	3	4	7	4	5	4	6	5	2
Mean excitement (SD)	3.79 (0.69)	5.26 (0.64)	5.37 (1.18)	4.79 (1.24)	5.58 (0.94)	4.74 (0.64)	5.16 (0.81)	4.74 (1.02)	3.21 (1.36)
Overall excitement	4	6	7	5	6	6	5	6	2
Difference	+0.21	+0.74	+1.63	+0.21	+0.42	+1.26	-0.16	+1.26	-1.21

a difference equal to 0. The participants with most differences of > 0.50 are P01, P08, P18 and P20 with 4 out of 5 instances in this category each, and P05, P12, P16 and 17 with 3 instances in the category.

In 7 out of 90 cases, the difference between a participant's overall assessment and their mean assessment through the experiment was ≥ 1.50 , which is indicated by the red numbers. In this case, we defined the category with 1.50 because anything below might as well indicate 1, and we assumed that to be just as likely to happen by chance as for inconsistency reasons considering we used a 7-point scale. At the same time,

APPENDIX 6. VALIDATION

Table 6.2: The table shows the general mood assessments indicated by *Initial [parameter]* and the mean assessment including standard deviation given by participant **P11 to P20** (except for P14) for each of the 5 parameters; *enjoyment, frustration, challenge, boredom, and excitement*. The table also shows the overall assessment that the participants gave for each of the parameters, which is compared to the mean for the same parameter resulting in a row indicating this difference. Differences of > 0.5 is indicated with green and differences of ≥ 1.5 is indicated with red.

	P11	P12	P13	P15	P16	P17	P18	P19	P20
Initial enjoyment	4	4	2	6	5	4	2	6	3
Mean enjoyment (SD)	3.68 (1.08)	3.74 (0.71)	4.84 (0.99)	5.00 (1.72)	5.11 (0.97)	5.53 (0.68)	5.11 (1.17)	5.37 (0.58)	4.89 (1.02)
Overall enjoyment	3	3	6	6	6	6	6	6	5
Difference	-0.68	-0.74	+1.16	+1.00	+0.89	+0.47	+0.89	+0.63	+0.11
Initial frustration	1	1	1	1	2	1	1	3	4
Mean frustration (SD)	1.16 (0.36)	1.00 (0.00)	1.16 (0.49)	2.26 (1.86)	3.00 (1.56)	1.11 (0.31)	2.05 (1.05)	2.89 (1.07)	3.11 (1.12)
Overall frustration	1	1	2	2	3	2	2	4	3
Difference	-0.16	0.00	+0.84	-0.26	0.00	+0.89	-0.05	+1.11	-0.11
Initial challenge	1	2	1	1	2	1	1	4	3
Mean challenge (SD)	1.63 (0.93)	1.16 (0.36)	2.11 (1.21)	3.11 (1.55)	3.11 (1.25)	1.16 (0.49)	3.89 (1.62)	4.05 (1.28)	3.16 (0.99)
Overall challenge	3	1	3	1	2	3	4	4	3
Difference	+1.37	-0.16	+0.89	-2.11	-1.11	+1.84	+0.11	-0.05	-0.16
Initial boredom	4	4	1	1	4	1	3	1	5
Mean boredom (SD)	4.42 (1.04)	4.42 (0.88)	1.16 (0.36)	1.42 (0.75)	4.05 (0.22)	1.00 (0.00)	2.05 (0.89)	2.21 (0.61)	3.78 (1.06)
Overall boredom	2	5	2	1	4	1	2	2	4
Difference	-2.42	+0.58	+0.84	-0.42	-0.05	0.00	-0.05	-0.21	+0.21
Initial excitement	4	4	5	7	6	5	6	7	3
Mean excitement (SD)	3.16 (1.09)	3.68 (0.86)	4.89 (0.64)	5.11 (1.07)	5.16 (0.67)	5.68 (0.46)	5.00 (0.65)	5.63 (0.67)	4.42 (0.94)
Overall excitement	2	4	5	4	5	5	5	7	5
Difference	-1.16	+0.32	+0.11	-1.11	-0.16	-0.68	0.00	+1.37	0.58

assuming that their overall assessment should be exactly equal to their mean assessment, would indicate that the participants attribute exactly equal meaning to each step. This is likely not the case, not only because the time they spent on each step varied greatly, but also because affective memory is not likely to be attributed exactly equally through a range of steps (Trakas, 2021). P02 had 2 out of 5 instances in this category, being the only participant with more than one instance in this category. All participants with instances in this category also had instances of > 0.50 for other parameters, meaning that there seems to be no pattern for the individual participants. The parameter with

most instances in this category was *challenge*, with 4, but there were 2 instances where the participants assessed it more challenging overall and 2 instances where they assessed it less challenging, indicating that there is no pattern to find here either.

Given that only 7 out of 90 instances in assessments were ≥ 1.5 , and 41 out of 90 instances in assessments were > 0.50 , this could indicate that the participants generally were able to use the scale consistently throughout the experiment. However, table 6.1 and 6.2 also show the general mood assessed by the participants before the experiment started. Here, it can be seen that the participants' general mood assessments either did not deviate at all or deviated 1 from their overall ranking in almost all instances where the difference between the mean rankings and the overall rankings were > 0.50 . There were only 4 instances, where this was not the case, two being for the parameters *frustration* and *boredom* for P02, who gave a general mood assessment of 4 and 5 and an overall assessment of 2 and 3, respectively. The third instance is for the parameter *enjoyment* for P17, who gave a general mood assessment of 4 but an overall assessment of 6, with the fourth being for the parameter *challenge* for P18, who gave a general mood assessment of 1 but an overall assessment of 4. This indicates that the participants' overall assessments are likely to have been influenced by their general mood before the experiment in many cases, and that their mood therefore was not affected much by the experimental setup.

6.2 Assessment and Debriefing Response Comparison

The second part of the validation was related to the questions in the debriefing part on the last page of the QS, where the participants were asked to mention moments that they enjoyed the most, as well as felt the most frustrated, challenged, bored, and excited. As these questions were meant to be used to validate if there was a match between the highest rated steps and the steps mentioned in the debriefing.

Even though the participants were not limited to having to mention a specific step in the debriefing, they often did mention one or several steps. Table 6.3 shows the cases, where the steps rated highest in the assessment and the steps mentioned in the debriefing were similar. Here, it is seen that there are most *partially similar* responses for all parameters except for *excitement*. *Partially similar* refers to either some, but

not all the steps with the highest rating being mentioned in the debriefing, or that some but not all steps mentioned in the debriefing were given the highest rating. The reason that there are so many *partially similar* instances could be that the participants were not limited to only mentioning one moment, which is probably why the participants mentioned up to four steps in the debriefing. Another reason is that in several cases, the participants gave more than one step the highest rating, and P17 actually gave 12 steps the highest rating for *enjoyment*, while only mentioning one of those steps in the debriefing, resulting in *partial similarity*. This could be an indication of the 7-point scale not being able to measure small enough differences, since some steps might have had the same rating, but one or two should have been slightly higher, and thus were more memorable regarding the specific parameter than others.

Table 6.3: The table shows how many times the participants mentioned a specific step, when asked to mention moments regarding the 5 different parameters in the debriefing. The table also shows the cases where the step given the highest assessment and the mentioned step in the debriefing were *partially similar*, as well as the cases where they were *completely similar*. Lastly, it shows the cases where at least one step given the highest rating was mentioned in the debriefing.

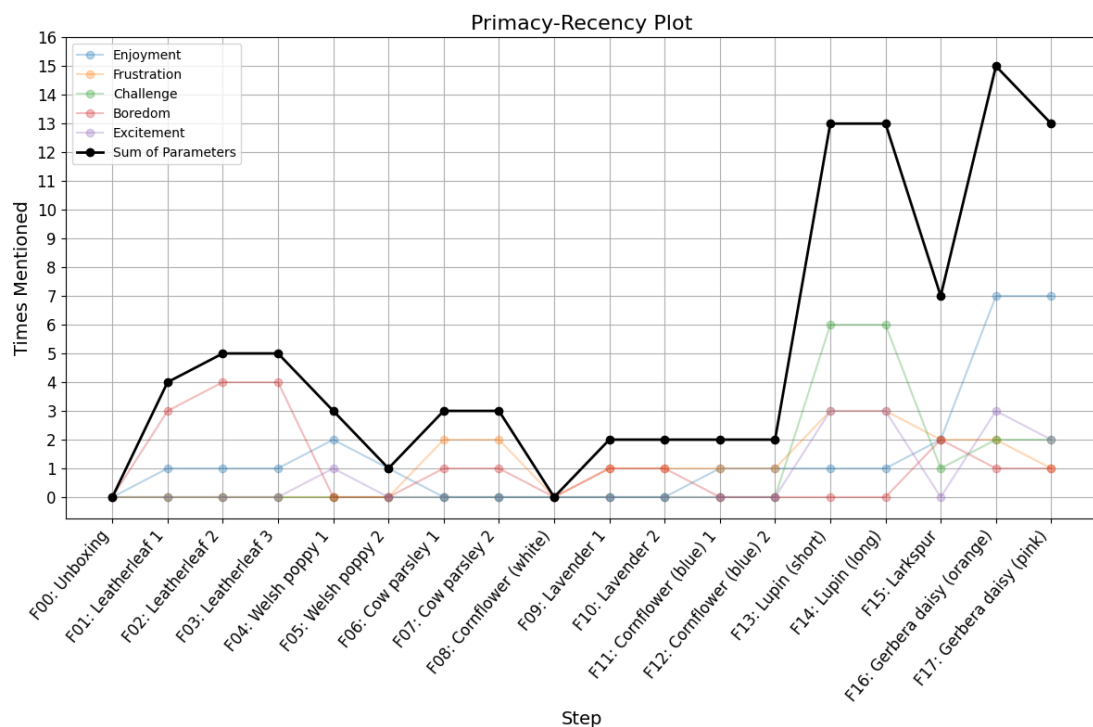
	Enjoyment	Frustration	Challenge	Boredom	Excitement
Specific step mentioned	11	9	7	3	7
Partially similar	9	6	4	3	2
Completely similar	1	1	2	0	2
Similarity sum	10	7	6	3	4

6.2.1 Primacy and Recency

We looked into the primacy and recency effect to investigate whether the specific steps mentioned by the participants during the debriefing were affected by these (Nordfang and Nørby, 2017, p. 152). Figure 6.1 shows how many times, each step was mentioned in total, as well as in relation to each of the 5 parameters. From the figure, we see that especially the recency effect may have affected the participants' responses during the debriefing. In relation to *enjoyment*, the last 2 steps were the ones mentioned most frequently, however, these steps were only mentioned one more time than F13 and F14. Regarding the primacy effect, we see an increase in mentions of the first 3 flowers (F1, F2, and F3) which were mentioned about 4 times, when asked about boredom. Looking

at the sum of mentions for each step, we do see a pattern that could indicate an effect of both the primacy effect and the recency effect, with the recency effect being the more prominent of the two. However, it is worth noting that the participants had all 17 flowers next to them while answering the debriefing questions, meaning that they had access to memory cues while answering the questions. While watching the videos, we found that the participants often showed the flower, they were mentioning during the debriefing, which means that they did use the physical flowers as cues during the debriefing.

Figure 6.1: The figure shows the frequency of mentions for each step for each of the 5 parameters. The sum of mentions for all steps can also be seen.



The results of this part of the validation indicate that participants answered the assessment scales consistently, since they in several cases recalled the same flowers as they gave the highest rating of the parameter they were asked to recall specific moments in connection with. Although, we see indications of especially the recency effect, we believe that the memory cues of the flowers being present, were sufficiently strong to diminish this effect, and that their responses were genuine.

7 Discussion

The aim of conducting the experiment was to develop a method for collecting UX metrics by analyzing facial action units (AU). This involved a remote data collection setup, as well as a foundation for training machine learning models to predict the 5 self-assessment parameters; *enjoyment*, *frustration*, *challenge*, *boredom*, and *excitement*. In this chapter, we will discuss the outcome and results of this experiment, as well as approaches to potential improvements for future work in the field.

The models' performance, as indicated by the reported metrics; Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 for the regression models, along with accuracy, precision, recall, and F1 score for the classification models, suggests a limited predictive capacity for the trained models. For instance, the reported regression model for *mean* data, yielded an R^2 value of between -2.78 and 0.06 (see table 5.2), indicating that the models explain virtually none of the variance in the ratings. A negative R^2 value indicates that the mean is as good of an indicator as the predicted value, which is inadequate (Chicco et al., 2021). Similarly, when we attempted to include all individual data points, the models performed equally unsatisfactory. The same can be seen for the classification models, that showed low accuracy, precision, recall, and F1 scores, both using mean as well individual data points. Through numerous attempts of adjusting data and hyperparameters, no successful models were achieved.

These results imply that the facial expression data, as collected and processed, may not sufficiently capture the nuances of the participants' emotional experiences to accurately predict their self-reported assessments. This could be due to several factors, that will be explained in the following.

7.1 Expression of emotions

When looking at Figure 5.9, we see that the AUs that were supposed to be active for *enjoyment* (AU6 and AU12), as explained in Appendix 3.3, do not seem to be affected by their experience of enjoyment, or at least their self-assessment hereof. The same applies for *excitement*, which can be seen on the same figure, where the activation of AUs were based on an assumption that *excitement* is connected with *enjoyment*. When looking at *frustration*, there seems to be no connection between any of the AUs found in the literature and the self-assessed rating of the parameter. This can be seen in Figure 5.10, where the AUs found in the literature (being all the AUs shown in the figure except AU6) seem to have no correlation to the rating of self-assessed *frustration*. For *boredom*, the literature gave us two options: either AU4, AU7 and AU12 were supposed to be active, or no AUs were supposed to be active. Here, Figure 5.10 shows that the activation of AU4, AU7 and AU12 seemed to be negatively affected by the experience of boredom, or at least the self-assessment hereof. This means that a high self-assessment of *boredom* seems to have resulted in a lower intensity of the 3 AUs than a low self-assessment. Before assigning too much importance to these graphs, one should note that the number of data points vary substantially between each assessment level, due to the fact that e.g. low ratings for *enjoyment* and *excitement* were rarely selected, and vice versa for *frustration*, *challenge* and *boredom*.

Overall, the missing correlation between the intensity of AUs and the rating of the parameters could imply that the parameters (excluding *challenge*) were not related to the emotions, and therefore the activation of AUs, in the way that was assumed. Another possibility is that the participants being alone during the experiment have resulted in a different activation of AUs than from the contexts researched in literature, where other people were present. This is supported by Fridlund (1991), who found that some facial expressions are socially dependent. To investigate this, future work could modify the experimental approach to focus on building as a social activity. A third explanation, that might be less likely, is that while the 6 basic emotions are considered to be expressed similarly across cultures, genders, and age groups, the 5 parameters researched in this study, could be subject to individual differences (Ekman and Friesen, 1978; Tejada et al., 2022; Sharma et al., 2022). An example observed in the videos was that

some participants seemed to express what we perceived as concentration through lowering their brow while others stuck the tip of their tongue out. This variability could be an explanation for the reduction in the models' ability to generalize across different individuals.

Not only the specific expression of emotions, but also the intensity of different AUs could be affected by individual differences. The decision to normalize AU intensity values both by the maximum value any AU can reach and by the individual participant's maximum value was a thoughtful attempt to address this variability in expressiveness. However, the fact that neither normalization approach significantly improved the models' performance suggests that the underlying data might still lack sufficient clarity or relevance for the task at hand. While switching to training the model on individual data points led to a substantially larger data set with more than 1 million data points, these still originated from a relatively small sample size, causing vulnerability to overfitting the models. An approach to limiting this vulnerability could be to include a larger sample size, and e.g. ask the participants to only build one or two flowers each.

Human emotions, particularly in an interactive and engaging activity like building with LEGO, are complex and influenced by various internal and external factors. The five parameters assessed may therefore not manifest in facial expressions alone or may do so in subtle ways that are difficult to capture consistently. This is supported by looking at Figure 5.3, where we see that there are rare instances of significant differences between the steps when looking at the parameters individually. Furthermore, from looking at Figure 5.4, there is no indication of a pattern emerging from the mean intensity of AUs over the course of the entire experiment. The results shown in the two figures imply that we did not succeed in affecting the participants with the task of the unboxing and building the flowers in the experiment. A reason for this could be that building with LEGO is not a task affecting the experiencing of the 5 parameters much, even though we chose to use a LEGO set targeted at adults in an attempt to give the participants a task that varied as much as possible. Moreover, Figure 5.2 shows that especially for the parameters *frustration*, *challenge* and *boredom*, the participants did not use the rating 7 very often, which is also the case for the rating 1 for *enjoyment* and *excitement*. This further indicates that the participants did not feel strongly affected by the task in the experiment. It is an issue that so few data points are available in the outer points, as this

leads the data set to be heavily skewed, which in turn affects the model performances negatively. Thus, a prediction of e.g. 1 in *enjoyment* might only be based on a single step for a single participant, which is in no way enough training data for a proper prediction, no matter if it is trained on the mean values or the individual data.

Thus, shifting from aggregated mean values to using individual data points, which theoretically should enhance the model's training effectiveness, only showed marginal improvements in the models' performance. This supports the former statement that merely increasing data points is not enough. Instead, each data point needs to robustly reflect the participant's emotional state, which might not be the case with the current dataset configuration, as we assign the rating given for the entire experience throughout one step to every single frame of their facial expression in that step. This means that we, possibly incorrectly, attribute high values of e.g. excitement to a step where the face actually has a neutral expression far more than half of the time spent on the step. This is probably what leads the model to have such poor performance, since we compromise high quality data of up to 15 minutes (see mean time spent on step on Figure 5.1) into a single assessment for each parameter. Thus, the poor performance of the models might not be due to AUs not reflecting the 5 parameters, but rather that the self-assessment should happen with smaller intervals between them. Initially, we considered it to be quite tedious if participants had to rate the 5 parameters possibly 100+ times, but if the aforementioned suggestion of asking participants to only build one flower each, multiple assessments might be achievable. For this, inspiration could be taken from Čertický et al. (2019), who asked participants to watch the video in retrospect and rate their selected parameter with specified intervals, depending on when their participants expressed a change in enjoyment.

7.2 Self-assessment Scale

As seen in Appendix 6, some participants used the same rating for more than half of the steps, indicating that the granularity of the scale was not fine enough. However, Lewis and Erdinç (2017) conducted an experiment comparing participants' use of a 7-point categorical scale, an 11-point categorical scale, and a visual analogue scale when answering the same questions, and found that there were no significant differences in

how participants assessed the three types of scales. Apart from this, the 7-point scale is the most commonly used in the literature regarding UX (Lewis and Erdiñç, 2017). Furthermore, as Lewis and Erdiñç also point out as a possible limitation of their study, our participants might have used their previous assessments as a guide to keep the following assessments consistent. This might have influenced our experiment as the participants had the entire question sheet (QS) in front of them during the entire experiment and were able to look at prior assessments, which might be why some participants tended to rate similar flowers the same. Of course, this might also be their genuine assessment, since it usually took less than 30 seconds to rate the assessments per flower, meaning that they probably did not review their prior assessments. Even so, they most likely remembered their prior ratings.

Using the prior responses as a guide can usually be avoided by clear unambiguous labels in the scale and definitions of the parameters. While we attempted to reach this by labelling the scales with the parameter as well as *Low*, *Neutral* and *High*, they might not have a clear understanding of how each of the parameters are defined. They might attribute several meanings to e.g. the word *excitement*, but not consider these consistently throughout all the steps in the experiment. Mayer et al. state that we have most likely come to a general understanding of the meaning of some emotions, but also point out that this does not mean that there is one correct way to interpret an emotion. Therefore, it is possible that the 5 parameters in the experiment do not elicit a general understanding among people, making them more difficult for the participants to assess. Apart from this, the participants were not given a point of reference for the assessment, which further means that they might have been unsure how to assess the parameters. One way to improve this could be to purposely add a guide or reference by making the participants build a reference flower, which they would then have to assess all other flowers compared to, hopefully making the assessments more consistent. This would have also minimized the opportunity for neutral and extreme response style biases affecting the assessments (Wetzel et al., 2016).

7.3 Experimental design

The experiment was carried out in English, despite all of the participants being native Danish speakers, which means that the participants were not able to complete the experiment in their native language, except for the debriefing part of the experiment where the participants read questions in English, but were allowed to answer in Danish. 5 participants out of the 18 that we received data from, chose to answer the questions in English. It is unknown whether this was because they did not realize that they were allowed to answer in Danish, or if they preferred to answer in English, e.g. because the rest of the experiment was in English. Even though all participants were fluent in English, it might have affected their assessments of the 5 parameters, since these were not translated into Danish. This became apparent as some participants used different words in Danish to e.g. describe *enjoyment* during the debriefing. This could have possibly been avoided had we given them a clear definition or supplied them with a Danish translation.

The experiment was carried out with a remote setup, as the only one of its kind, meaning that the completion of the experiment relied on easy-to-follow instructions, making it clear what the participants had to do. In our case, this resulted in 2 participants not being able to send their recording back to us due to a malfunction in their recording device. Other than this, only small issues occurred such as participants forgetting to fill out the QS right after building a flower, requiring assistance for sending the material back to us, and not hearing from some participants until far beyond the set deadline. Overall, this method was thought to be the best fit for the experiment in this study, since it increased the ecological validity, was less time consuming, because we did not have to be present for the data collection, and finally, it gave the participants the freedom to complete the experiment, when they had the time.

Regarding the ecological validity, it is worth noting that the setting of the experiment was quite unnatural for some of the participants as they were not allowed to listen to music or watch TV and had to be quiet and alone while building. Three participants pointed out during the debriefing that they would have liked to listen to music, watch TV, or talk to someone while building, if they had been allowed to do so. One of these participants even pointed out during the debriefing that they thought of LEGO building as a social activity. The remaining two participants stated that this made it slightly

boring to participate in the experiment. This is also evident in the responses from the recruitment survey, where 7 participants indicated that they build with LEGO as part of a social activity or for the sake of others. Five other participants also pointed out that they thought the experiment was a bit boring, and even though most of them stated that it was because of the building being repetitive, there is a chance that the setting of the experiment might have affected these participants' engagement as well. As described in Appendix 2, engagement was measured by Holiday et al. (2023) and Rodríguez-Fuertes et al. (2022), and it might have been beneficial to also measure it in our experiment to gain further knowledge on this topic. It is worth noting, however, that we decided to focus on few parameters during the experiment, as we did not want the participants to have to answer too many questions, as all the parameters were assessed 20 times in total during the experiment, which was already quite time consuming and tedious to complete.

Research has shown that seeing a video feed of yourself, e.g. while in an online video call, can be distracting and lead to self-awareness, which led us to consider the fact that the participants were filmed during the experiment could have affected their facial expressions (Qu et al., 2017). In many cases, the participants set up to their recording device, so that they could see themselves meanwhile. However, this way of conducting the experiment was assumed to be less intrusive compared to e.g. a direct observation and an interview. Some participants pointed out that they were very aware of the recording in the beginning of the experiment, but forgot it as they proceeded. A way to limit this bias was to adapt the method described by Parvanta et al. (2022). In their experiment, an online survey was sent out consisting of videos and questions regarding those videos, where the platform was set up to record the subjects whenever they watched one of the videos to collect data on their facial expression. An issue with implementing this in our experiment is that we used a tangible product, which would have made it difficult to set up a platform to record when the participants were building. One way to do it, however, would be to develop a platform including the QS and make it record every time they had assessed the parameters and clicked next to move onto the next step. This way, the platform could record until the participants indicated that they had finished the given step, after which the assessment parameters would be shown again. This would, however, make the experiment setup more vulnerable to participant

errors in e.g. forgetting to answer the QS between steps, and might lead to more data being lost. It would also have required more effort to develop and set up, but might have made the data collection easier and limited the bias of the participants being aware that they were being recorded. Either way, the participants would have known that they were being recorded, and there is no way of knowing if this solution would have made them less aware of this.

The order, the 5 parameters were shown to the participants, were balanced between participants. It could have also been beneficial to balance the order of the flowers in the LEGO set in order to reduce carry-over effects. Furthermore, 12 participants pointed out that the experiment was quite repetitive because they had to build some flowers more than once and right after one another. One participant suggested changing the order of the flowers, since this would have made the experiment feel less repetitive and thus more interesting. This is, however, an issue regarding the order of the LEGO manual, since we did not change this in the experiment. The instruction manual for set 10313 is available through the LEGO website, meaning that we could have downloaded it, changed the order of the flowers and told the participants that they had to follow a different version of the manual that we would hand them instead of the instruction manual found in the LEGO box. A downside to this approach is, however, that the pieces for the flowers are separated into bags that follow the order of the instruction manual, meaning that the order could not easily have been balanced between all flowers, but instead the flowers in one bag could have been counterbalanced and the order that the participants had to open the bags could also have been counterbalanced. One could argue, that this would have decreased the ecological validity, since the LEGO set would most likely be built by following the real instruction manual exactly by the people who would buy or receive the set.

7.4 Future work

The findings highlight the necessity for further refinement in both data collection and model training approaches. With the priorly described enhancements to the data collection method, exploring advanced machine learning and deep learning methods might also be beneficial. These can capture complex patterns and interactions in the data, and

might enhance predictive performance. Techniques like convolutional neural networks (CNNs) for image analysis or recurrent neural networks (RNNs) for time-series data could be particularly useful (Bishop, 2006, p. 267-269; Petneházi, 2019). Kremsner et al. (2023) point out that only single images are evaluated in facial expression analysis, which means that dynamic changes in facial expression are not evaluated. While we attempted to introduce dynamics into the data set, it might not have been sufficient, hence why RNNs could be a useful next step. By using RNNs, the model could take the temporal aspect into account, allowing each step to be interpreted as part of a time sequence. This approach would enable the analysis of the activation of AUs in different sequences and help identify patterns within each step and across steps for certain subjective assessments. However, with the implementation of neural networks, the interpretability of the model is substantially reduced, meaning we would have no way of interpreting which AUs are affecting the participants' experiences.

Assuming the success of implementing the refined methodological approach and model, a real-life application of the research done in this study involves applying it in real-time user testing. This would allow UX researchers to probe participants by pointing out when certain parameters are especially active/not active and ask for the them to explain their thoughts and reasoning. In a context related to building with LEGO, this means that it would be possible to point out a specific step in the instruction manual, where participants are more emotionally affected, e.g. because a lot of mistakes are made in this step. In this study, we found that the participants were not necessarily affected by making a mistake right away, because they did not realize that a mistake was made, however, when they noticed, it became evident in their facial expressions in the video. This scenario would still make the use of the proposed method relevant, since the participants would be able to track back to the exact step where the mistake was made. This way of using the method would also mean that the researchers would not have to watch an entire video of an experiment afterwards, making the data collection and the analysis more efficient and less time consuming.

A way of validating a model based on data collected in a study like this, would be by comparing the data with biometric data such as electroencephalogram (EEG) or galvanic skin response (GSR). These methods are often seen used in combination with FEA (Santoso et al., 2016; Čertický et al., 2019; Bañuelos-Lozoya et al., 2021; Brunken et

al., 2003). One reason for not including them in the current study was that these sensors have to be attached to the participants' head or fingers, which is not ideal when having to use your fingers for building LEGO. EEG sensors are very sensitive and require a controlled setup, practically impossible for a remote study. Furthermore, both methods are seemingly more invasive than having your face recorded, also leading to decreased ecological validity. However, it might have still been relevant to include some sort of separate biometric validation that would disturb the participants as little as possible, given that Parvanta et al. (2022) state that it can be difficult to estimate the affective state of participants in a non-lab setting without other physiological sensors than facial expression recognition software. This is because many factors can affect a person's facial movements, such as lighting facing them causing their brows to lower and eyes to squint, and thus activating AU4 and AU7.

8 Conclusion

The purpose of this study was to collect data in a remote setting suitable for evaluation of user experience (UX) using facial expression analysis (FEA). The video material received was suitable for FEA, however, the second objective of the study, being to train models, leveraging action units (AUs) identified by the FEA, to predict participants' subjective assessments of the five parameters: enjoyment, frustration, challenge, boredom, and excitement was less successful.

Prior to conducting an experiment, a literature review was carried out to investigate the use of FEA in literature and to gain inspiration for our study. Here, it was found that FEA is often used in relation to subjective assessments in e.g. sensory tests, but rarely in UX contexts. Furthermore, FEA was rarely used in remote setups. The experiment focused on collecting data on facial expressions as well as self-assessed rankings of the parameters: enjoyment, frustration, challenge, boredom and excitement on a 7-point scale ranging from low to high. This was done through a remote experimental setup, including 18 participants who were asked to record themselves while building a LEGO set consisting of 18 steps.

The video data received from the participants was run through the OpenFace 2.0 software to analyze the presence and intensity of so-called facial action units (AU) for each participant for all 18 steps. Machine learning models, based on random forest regression and classification, were trained to predict the assessment of the 5 parameters based on the activation of AUs. In spite of several approaches to improve the models, e.g. through data formatting and normalization and hyperparameter adjustments, performance metrics still showed poor predictability of the subjective parameters.

For future research, we propose several adjustments to improve the outcome of the experiment and the analysis. Regarding the method, it would be beneficial to collect

self-assessment data more frequently, since only one assessment was given per parameter for each of the 18 steps, where some steps took more than 10 minutes to complete. To avoid asking single participants to assess the same parameters 100+ times, sample size should also be increased, which in turn would improve the generalizability of the models. Moreover, applying more complex approaches for training the models should also be attempted, as the presented ones do not consider time-series data.

In conclusion, the method from this study was suitable for remote data collection to be analyzed with FEA, and to be used to train machine learning models. However, the collection of self-assessments, the target of our models, needs further refinement, before satisfactory results can be achieved. Although the current models show limited predictive capability, they do provide a foundation for ongoing research into the intricate relationship between facial expressions and subjective emotional experiences. This along with the suggested improvements of the methodology and analysis warrants space for future research.

Bibliography

- Abler, B., Walter, H., & Erk, S. (2005). Neural correlates of frustration. *NeuroReport*, 16(7). <https://doi.org/10.1097/00001756-200505120-00003>
- Baltrušaitis, T. (2019). Command line arguments [Accessed: May 24 2024]. <https://github.com/TadasBaltrušaitis/OpenFace/wiki/Command-line-arguments>
- Baltrušaitis, T. (2021). Mac installation [Accessed: May 24 2024]. <https://github.com/TadasBaltrušaitis/OpenFace/wiki/Mac-Installation>
- Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. <https://doi.org/10.1109/WACV.2016.7477553>
- Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. <https://doi.org/10.1109/FG.2018.00019>
- Bañuelos-Lozoya, E., González-Serna, G., González-Franco, N., Fragoso-Díaz, O., & Castro-Sánchez, N. (2021). A systematic review for cognitive state-based qoe/ux evaluation. *Sensors (Basel, Switzerland)*, 21(10), 3439–. <https://doi.org/10.3390/s21103439>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2), 281–305.
- Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(1), 1063–1095. <https://doi.org/10.48550/arXiv.1005.0208>
- Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

- Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 53–61. https://doi.org/10.1207/S15326985EP3801_7
- Carmichael, L., Poirier, S.-M., Coursaris, C., Léger, P.-M., Sénécal, S., Davis, F. D., Riedl, R., vom Brocke, J., Léger, P.-M., Randolph, A. B., & Müller-Putz, G. (2021). *Does media richness influence the user experience of chatbots: A pilot study* (Vol. 52). Springer, https://doi.org/10.1007/978-3-030-88900-5_23
- Ceccacci, S., Generosi, A., Giraldi, L., & Mengoni, M. (2023). Emotional valence from facial expression as an experience audit tool: An empirical study in the context of opera performance. *Sensors (Basel, Switzerland)*, 23(5), 2688–. <https://doi.org/10.3390/s23052688>
- Čertický, M., Čertický, M., Sinčák, P., Magyar, G., Vaščák, J., & Cavallo, F. (2019). Psychophysiological indicators for modeling user experience in interactive digital entertainment. *Sensors (Basel, Switzerland)*, 19(5), 989–. <https://doi.org/10.3390/s19050989>
- Chang, D., Yin, Y., Li, Z., Tran, M., & Soleymani, M. (2024a). Libreface: An open-source toolkit for deep facial expression analysis. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8190–8200. <https://doi.org/10.48550/arXiv.2308.10713>
- Chang, D., Yin, Y., Li, Z., Tran, M., & Soleymani, M. (2024b). Libreface: An open-source toolkit for deep facial expression analysis [Accessed: May 20 2024]. <https://github.com/ihp-lab/LibreFace/tree/main>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ. Computer science*, 7, 623–623. <https://doi.org/10.7717/peerj-cs.623>
- Clark, E. A., Duncan, S. E., Hamilton, L. M., Bell, M. A., Lahne, J., Gallagher, D. L., & O’Keefe, S. F. (2021). Characterizing consumer emotional response to milk packaging guides packaging material selection. *Food quality and preference*, 87, 103984–. <https://doi.org/10.1016/j.foodqual.2020.103984>
- Cohn, J., Schmidt, K., Gross, R., & Ekman, P. (2002). Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to in-

BIBLIOGRAPHY

- form person identification. *International Conference on Multimodal Interfaces: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces; 14-16 Oct. 2002*, 491–496. <https://doi.org/10.1109/ICMI.2002.1167045>
- Crist, C., Duncan, S., Arnade, E., Leitch, K., O’Keefe, S., & Gallagher, D. (2018). Automated facial expression analysis for emotional responsivity using an aqueous bitter model. *Food quality and preference*, 68, 349–359. <https://doi.org/10.1016/j.foodqual.2018.04.004>
- Cutler, R. D., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792. <https://doi.org/10.1890/07-0539.1>
- Deng, W., Jia, M., & Zhang, Z. (2023). How corporate social responsibility moderates the relationship between distributive unfairness and organizational revenge: A deontic justice perspective. *Chinese management studies*, 17(6), 1240–1258. <https://doi.org/10.1108/CMS-09-2021-0400>
- de Wijk, R. A., Ushijima, S., Ummels, M., Zimmerman, P., Kaneko, D., & Vingerhoeds, M. H. (2021). Reading food experiences from the face: Effects of familiarity and branding of soy sauce on facial expressions and video-based rppg heart rate. *Foods*, 10(6), 1345–. <https://doi.org/10.3390/foods10061345>
- Ekman, P. (1992). Facial expressions of emotion: New findings, new questions. *Psychological science*, 3(1), 34–38. <https://doi.org/10.1111/j.1467-9280.1992.tb00253.x>
- Ekman, P., & Friesen, W. V. (1978). Facial action coding system (facs). *APA PsycTests*. <https://doi.org/10.1037/t27734-000>
- Ekman, P., & Rosenberg, E. (2005). *What the face reveals : Basic and applied studies of spontaneous expression using the facial action coding system (facs)* (2nd Edition). Oxford University Press.
- Ferrari, R. (2015). Writing narrative style literature reviews. *Medical writing (Leeds)*, 24(4), 230–235. <https://doi.org/10.1179/2047480615Z.000000000329>
- Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of personality and social psychology*, 60(2), 229–240. <https://doi.org/10.1037/0022-3514.60.2.229>

- Gosselin, P., Perron, M., Legault, M., & Campanella, P. (2002). Children's and adults' knowledge of the distinction between enjoyment and nonenjoyment smiles. *Journal of nonverbal behavior*, 26(2), 83–108. <https://doi.org/10.1023/A:1015613504532>
- Gülşen, M., Aydın, B., Güreş, G., & Yalçın, S. S. (2023). Ai-assisted emotion analysis during complementary feeding in infants aged 6–11 months. *Computers in biology and medicine*, 166, 107482–107482. <https://doi.org/10.1016/j.combiomed.2023.107482>
- Halamová, J., Kanovský, M., Brockington, G., & Strnádelová, B. (2023). Automated facial expression analysis of participants self-criticising via the two-chair technique: Exploring facial behavioral markers of self-criticism. *Frontiers in psychology*, 14, 1138916–1138916. <https://doi.org/10.3389/fpsyg.2023.1138916>
- Hammond, R. W., Parvanta, C., & Zemen, R. (2022). Caught in the act: Detecting respondent deceit and disinterest in on-line surveys: a case study using facial expression analysis. *Social marketing quarterly*, 28(1), 57–77. <https://doi.org/10.1177/15245004221074403>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
- Holiday, S., Hayes, J. L., Park, H., Lyu, Y., & Zhou, Y. (2023). A multimodal emotion perspective on social media influencer marketing: The effectiveness of influencer emotions, network size, and branding on consumer brand engagement using facial expression and linguistic analysis. *Journal of interactive marketing*, 58(4), 414–439. <https://doi.org/10.1177/10949968231171104>
- Huntington, C. (2024). Frustration: Definition, examples, & principles [Accessed: May 21 2024]. <https://www.berkeleywellbeing.com/frustration.html>
- Janitza, S., Tutz, G., & Boulesteix, A.-L. (2016). Random forest for ordinal responses: Prediction and variable selection. *Computational statistics & data analysis*, 96, 57–73. <https://doi.org/10.1016/j.csda.2015.10.005>
- Jones, T., Randolph, A. B., Sneha, S., Davis, F. D., Riedl, R., vom Brocke, J., Léger, P.-M., Randolph, A. B., & Müller-Putz, G. (2021). *Examining the impact of social video game tournaments on gamers' mental well-being* (Vol. 52). Springer, https://doi.org/10.1007/978-3-030-88900-5_20

- Kassas, B., Palma, M. A., & Porter, M. (2022). Happy to take some risk: Estimating the effect of induced emotions on risk preferences. *Journal of economic psychology*, 91, 102527–. <https://doi.org/10.1016/j.joep.2022.102527>
- Krause, P. A., Kay, C. A., & Kawamoto, A. H. (2020). Automatic motion tracking of lips using digital video and openface 2.0. *Laboratory phonology*, 11(1). <https://doi.org/10.5334/labphon.232>
- Kremsner, T. P., Pfeiffer, C., Weidinger, S., & Stolavetz, C. (2023). How to visualize electricity consumption anomalies: The impact of chart types on triggered emotions and eye movements. *e-Prime*, 5, 100202–. <https://doi.org/10.1016/j.prime.2023.100202>
- Lawrence, K., Campbell, R., & Skuse, D. (2015). Age, gender, and puberty influence the development of facial emotion recognition. *Frontiers in psychology*, 6, 761–761. <https://doi.org/10.3389/fpsyg.2015.00761>
- Lewis, J. R., & Erdiñç, O. (2017). User experience rating scales with 7, 11, or 101 points: Does it matter? *Journal of Usability Studies*, 12(2).
- Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PloS one*, 13(8), e0199661–e0199661. <https://doi.org/10.1371/journal.pone.0199661>
- López-Mas, L., Claret, A., Bermúdez, A., Llauger, M., & Guerrero, L. (2022). Co-creation with consumers for packaging design validated through implicit and explicit methods: Exploratory effect of visual and textual attributes. *Foods*, 11(9), 1183–. <https://doi.org/10.3390/foods11091183>
- Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology*, 25(2), 141–158. <https://doi.org/10.1080/01449290500331156>
- Matsufuji, Y., Ueji, K., & Yamamoto, T. (2023). Predicting perceived hedonic ratings through facial expressions of different drinks. *Foods*, 12(18), 3490–. <https://doi.org/10.3390/foods12183490>
- Mavromoustakos-Blom, P., Kosa, M., Bakkes, S., & Spronck, P. (2021). Correlating facial expressions and subjective player experiences in competitive hearthstone.

- ACM International Conference Proceeding Series*. <https://doi.org/10.1145/3472538.3472577>
- Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International journal of remote sensing*, 39(9), 2784–2817. <https://doi.org/10.1080/01431161.2018.1433343>
- Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion (Washington, D.C.)*, 1(3), 232–242. <https://doi.org/10.1037/1528-3542.1.3.232>
- McDaniel, B., D’Mello, S., King, B., Chipman, P., Tapp, K., & Graesser, A. (2007). Facial features for affective state detection in learning environments. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 29(29). <https://doi.org/10.1145/3385209>
- Mookherjee, S., Lee, J. J., & Sung, B. (2021). Multichannel presence, boon or curse?: A comparison in price, loyalty, regret, and disappointment. *Journal of business research*, 132, 429–440. <https://doi.org/10.1016/j.jbusres.2021.04.041>
- Nordfang, M., & Nørby, S. (2017). *Kognitionsspsykologi, 1th edition*. Samfundslitteratur.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *The American psychologist*, 17(11), 776–783. <https://doi.org/10.1037/h0043424>
- Orne, M. T. (2002). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *Prevention & treatment*, 5(1). <https://doi.org/10.1037/1522-3736.5.1.535a>
- Parvanta, C., Hammond, R., He, W., Zemen, R., Boddupalli, S., Walker, K., Chen, H., & Harner, R. (2022). Face value: Remote facial expression analysis adds predictive power to perceived effectiveness for selecting anti-tobacco psas. *Journal of health communication*, 27(5), 281–291. <https://doi.org/10.1080/10810730.2022.2100016>
- Paul Ekman Group. (2024). Micro expressions [Accessed: May 16 2024]. <https://www.paulekman.com/resources/micro-expressions/>

- Perron, M., & Roy-Charland, A. (2013). Analysis of eye movements in the judgment of enjoyment and non-enjoyment smiles. *Frontiers in psychology*, 4, 659–659. <https://doi.org/10.3389/fpsyg.2013.00659>
- Petneházi, G. (2019). Recurrent neural networks for time series forecasting. *arXiv.org*. <https://doi.org/10.1016/j.ijforecast.2020.06.008>
- Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301. <https://doi.org/10.1002/widm.1301>
- Qu, F., Yan, W.-J., Chen, Y.-H., Li, K., Zhang, H., & Fu, X. (2017). “you should have seen the look on your face...”: Self-awareness of facial expressions. *Frontiers in Psychology*, 8, 259679. <https://doi.org/10.3389/fpsyg.2017.00832>
- Rao, S., Wirjopawiro, S., Pons Rodriguez, G., Röggla, T., Cesar, P., & El Ali, A. (2023). Affective driver-pedestrian interaction: Exploring driver affective responses toward pedestrian crossing actions using camera and physiological sensors. *ACM International Conference Proceeding Series*, 300–310. <https://doi.org/10.1145/3580585.3607168>
- Richlan, F., Thürmer, J. L., Braid, J., Kastner, P., & Leitner, M. C. (2023). Subjective experience, self-efficacy, and motivation of professional football referees during the covid-19 pandemic. *Humanities & social sciences communications*, 10(1), 215–215. <https://doi.org/10.1057/s41599-023-01720-z>
- Rodríguez-Fuertes, A., Alard-Josemaría, J., & Sandubete, J. E. (2022). Measuring the candidates’ emotions in political debates based on facial expression recognition techniques. *Frontiers in psychology*, 13, 785453–785453. <https://doi.org/10.3389/fpsyg.2022.785453>
- Roy, A., Sénécal, S., Léger, P.-M., Demolin, B., Bigras, É., & Gagne, J. (2020). Measuring users’ psychophysiological experience in non-linear omnichannel environment. In *Hci international 2020 - late breaking papers: User experience design and case studies* (pp. 762–779, Vol. 12423). Springer International Publishing. https://doi.org/10.1007/978-3-030-60114-0_50
- Samant, S. S., Chapko, M. J., & Seo, H.-S. (2017). Predicting consumer liking and preference based on emotional responses and sensory perception: A study with

- basic taste solutions. *Food research international*, 100(Pt 1), 325–334. <https://doi.org/10.1016/j.foodres.2017.07.021>
- Samant, S. S., & Seo, H.-S. (2020). Influences of sensory attribute intensity, emotional responses, and non-sensory factors on purchase intent toward mixed-vegetable juice products under informed tasting condition. *Food research international*, 132, 109095–109095. <https://doi.org/10.1016/j.foodres.2020.109095>
- Santoso, H., Schrepp, M., Kartono Isal, R. Y., Yudha Utom, A., & Priyogi, B. (2016). Measuring the user experience. *The journal of educators online*, 13(1). <https://doi.org/10.9743/JEO.2016.1.5>
- Sass, J., & Fekete, L. V. (2022). Secrets revealed by boredom: Detecting and tackling barriers to student engagement. *International Conference on Advanced Learning Technologies (ICALT)*, 417–419. <https://doi.org/10.1109/ICALT55010.2022.00129>
- Savela-Huovinen, U., Toom, A., Knaapila, A., & Muukkonen, H. (2021). Sensory professionals' perspective on the possibilities of using facial expression analysis in sensory and consumer research. *Food science & nutrition*, 9(8), 4254–4265. <https://doi.org/10.1002/fsn3.2393>
- Schmidt, K. L., Cohn, J. F., & Tian, Y. (2003). Signal characteristics of spontaneous facial expressions: Automatic movement in solitary and social smiles. *Biological psychology*, 65(1), 49–66. [https://doi.org/10.1016/S0301-0511\(03\)00098-X](https://doi.org/10.1016/S0301-0511(03)00098-X)
- Scikit-learn. (2024a). *Randomforestclassifier* [Accessed: May 27 2024]. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- Scikit-learn. (2024b). *Randomforestregressor* [Accessed: May 27 2024]. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
- Scikit-learn. (2024c). *Randomizedsearchcv* [Accessed: May 27 2024]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
- Shah, M., Cooper, D. G., Cao, H., Gur, R. C., Nenkova, A., & Verma, R. (2013). Action unit models of facial expression of emotion in the presence of speech. 2013

- Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013*, 49–54. <https://doi.org/10.1109/ACII.2013.15>
- Sharma, K., Papavlasopoulou, S., & Giannakos, M. (2022). Children's facial expressions during collaborative coding: Objective versus subjective performances. <https://doi.org/10.1016/j.ijcci.2022.100536>
- Singh, D., & Singh, B. (2019). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 105524. <https://doi.org/10.1016/j.asoc.2019.105524>
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- Tejada, J., Freitag, R. M. K., Pinheiro, B. F. M., Cardoso, P. B., Souza, V. R. A., & Silva, L. S. (2022). Building and validation of a set of facial expression images to detect emotions: A transcultural study. *Psychological research*, 86(6), 1996–2006. <https://doi.org/10.1007/s00426-021-01605-3>
- The LEGO Group. (2024a). Adventures with mario starter course [Accessed: March 12 2024]. <https://www.lego.com/en-dk/product/adventures-with-mario-starter-course-71360>
- The LEGO Group. (2024b). The lego group history [Accessed: March 8 2024]. <https://www.lego.com/en-us/aboutus/lego-group/the-lego-group-history>
- The LEGO Group. (2024c). Wildflower bouquet [Accessed: March 12 2024]. <https://www.lego.com/en-dk/product/wildflower-bouquet-10313>
- Trakas, M. (2021). No trace beyond their name? affective memories, a forgotten concept. *L'année psychologique*, 121(2), 129–173. <https://doi.org/10.3917/anpsy1.212.0129>
- Walsh, A. M., Duncan, S. E., Bell, M. A., O'Keefe, S. F., & Gallagher, D. L. (2017). Breakfast meals and emotions: Implicit and explicit assessment of the visual experience. *Journal of sensory studies*, 32(3). <https://doi.org/10.1111/joss.12265>
- Wetzel, E., Bohnke, J. R., & Brown, A. (2016). Response biases. <https://doi.org/10.1093/med:psych/9780199356942.003.0024>

BIBLIOGRAPHY

- Wong, A., Wang, L., & Lee, J. R. (2020). Emotionnet nano: An efficient deep convolutional neural network design for real-time facial expression recognition. *Frontiers in Robotics and AI*, 7, 1–13. <https://doi.org/10.3389/frobt.2020.00001>
- Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2. <https://doi.org/10.1007/s41664-018-0068-2>
- Zarei, S. A., Yahyavi, S.-S., Salehi, I., Kazemiha, M., Kamali, A.-M., & Nami, M. (2022). Toward reanimating the laughter-involved large-scale brain networks to alleviate affective symptoms. *Brain and behavior*, 12(7), e2640–n/a. <https://doi.org/10.1002/brb3.2640>
- Zel, S., Duman, G., & Kongar, E. (2021). Improving online learning experience using facial expression analysis. *IEEE engineering management review*, 49(3), 71–81. <https://doi.org/10.1109/EMR.2021.3079840>
- Zeng, I. M., & Lobo Marques, J. A. (2023). Neuromarketing as a tool to measure and evaluate the consumer behaviour of guandong teahouse's social media advertisement. *ACM International Conference Proceeding Series*, 63–69. <https://doi.org/10.1145/3616712.3616787>
- Zhi, R., Hu, X., Wang, C., & Liu, S. (2020). Development of a direct mapping model between hedonic rating and facial responses by dynamic facial expression representation. *Food research international*, 137, 109411–109411. <https://doi.org/10.1016/j.foodres.2020.109411>
- Zhi, R., Wan, J., Zhang, D., & Li, W. (2018). Correlation between hedonic liking and facial expression measurement using dynamic affective response representation. *Food research international*, 108, 237–245. <https://doi.org/10.1016/j.foodres.2018.03.042>