Using games as educational tools:
An evaluation of a game for children to train facial expression recognition

EWOUT VAN DER HEIDE
Using games as educational tools: 
An evaluation of a game for children to train facial expression recognition

_Ett använda spel i utbildningssyfte:_
En utvärdering av ett spel riktat till barn för att träna förmågan att identifiera ansiktsuttryck

Ewoud van der Heide
Degree Project in Computer Science and Communication, Second Cycle 30 credits
Masters Programme in Human-Computer Interaction, 120 credits
KTH Royal Institute of Technology

_Supervisor: Christopher Peters_
_Examiner: Robert Comber_

June 7, 2018
Abstract

Facial expressions play a large role in non verbal communication. Research shows promising results for using games to improve facial expression recognition in children with autism spectrum disorder. Games are effective educational tools and are successful in motivating students. Using a game to improve facial expression recognition could be beneficial for all children as it reduces the risk for problematic behavior and mental health issues.

For this study a game to train facial expression recognition to children was developed and evaluated. The goal of the evaluation was to determine which factors influence performance and engagement in the game and if there are expressions that are often identified incorrectly. Additionally the children’s attitude towards the game was evaluated.

The results show that performance is affected by the difficulty, context and intensity. The children that showed the most engagement also performed better in the beginning of the game, however the correlation between performance and engagement is complex. Unfortunately it was not possible to evaluate the effect of rewards on the children’s engagement, but children were generally positive on the rewards. The confusion of expressions was in line with earlier research, but not as symmetrical. The players were generally positive about the game. Further research is needed to determine the long term learning effects of the game and to assess ways to engage players more.

Sammanfattning


Syftet med denna studie var att utveckla och utvärdera ett spel för barn för att träna förmågan att identifiera ansiktsuttryck. Målet med utvärderingen var att avgöra vilka faktorer som påverkar prestation och engagemang i spelet samt huruvida det finns uttryck som ofta identifieras inkorrekt. Ett ytterligare mål var att utvärdera barnens åsikter om spelet. 54 barn i åldrarna 8 och 11 år testade spelet vid två tillfällen.

ABSTRACT

Facial expressions play a large role in non verbal communication. Research shows promising results for using games to improve facial expression recognition in children with autism spectrum disorder. Games are effective educational tools and are successful in motivating students. Using a game to improve facial expression recognition could be beneficial for all children as it reduces the risk for problematic behavior and mental health issues.

For this study a game to train facial expression recognition to children was developed and evaluated. The goal of the evaluation was to determine which factors influence performance and engagement in the game and if there are expressions that are often identified incorrectly. Additionally the children’s attitude towards the game was evaluated.

The results show that performance is affected by the difficulty, context and intensity. The children that showed the most engagement also performed better in the beginning of the game, however the correlation between performance and engagement is complex. Unfortunately it was not possible to evaluate the effect of rewards on the children’s performance, but children were generally positive on the rewards. The confusion of expressions was in line with earlier research, but not as symmetrical. The players were generally positive about the game. Further research is needed to determine the long term learning effects of the game and to assess ways to engage players more.

Author Keywords
Facial expression, Emotion, Game, Children, Education

INTRODUCTION

"It is unfathomable that this emotional education is not a part of education at schools for everyone."

Written in an evaluation of a course aimed at partners of patients with emotional disorders

Currently, teaching people how to identify, effectively communicate and validate their own and others’ emotions is part of psychotherapy, to treat emotional disorders. Patients are often wondering why they were not taught these skills earlier in life, and later in therapy express frustration that other people are lacking them.
ment while playing the game. The factors that are evaluated with regards to performance are: the intensity of the shown expression, the context and the number of options to choose the correct expression from. The factors that are expected to affect engagement are: rewards for playing and the children’s performance. The analysis includes a comparison of the performance per emotion and whether specific emotions were often confused. Additionally, the children’s attitude towards the game is analyzed. The results are compared to earlier studies and discussed.

BACKGROUND
Emotion research is a divided field with many different theories [29]. Darwin [10] theorized that certain expressions of emotion are innate, meaning that they are universal and not learned. As proof he mentions that newborn infants show expressions of discomfort that are similar to older humans. Even people that have been born blind communicate their emotions with facial expressions that are similar to non-blind people. Tomkins [41] identified nine basic affects: interest, joy, surprise, distress, fear, shame, contempt, disgust and anger. However, up to this point the theories were based on observations in western society or cultures that had been exposed to western society. In 1971 Ekman [15] published a study to prove that there are six universal basic emotions: happiness, anger, sadness, disgust, surprise and fear. Stories of different emotional situations were told and subjects had to pick a photograph of a facial expression showing the matching emotion. He conducted the study with a tribe in New Guinea that had very little contact with western society. The study showed that with the exception of surprise, the emotions were judged in the same way as in western culture. Later studies confirmed these findings [16]. However, not all expressions are equally easily recognized; people are generally good at recognizing happy and angry, but confuse surprised and afraid [9].

Emotion recognition is influenced by the context the emotions is presented in [4]. For example, a marathon runner might cry after crossing the finishing line. If one would only look at the facial expression of the runner, the emotion will be interpreted differently than when looking at a picture showing more of the surroundings. Crying is usually associated with intense sadness, but in the case of the marathon runner it is more likely to express intense joy, relieve and exhaustion. Factors that influence the perception of emotion are: the description of the social situation, voice, posture, visual scenes and the personal background of the perceiver [6].

Impairment in emotion recognition
People with Autism Spectrum Disorder (ASD) can have problems with social-emotional communication; this can manifest as difficulty with non-verbal communication such as facial expressions [2]. It has been shown that young children with impaired emotion recognition skills show more behavioral problems [22] and have a higher risk of developing mental disorders [39]. Theory of mind is a term used to describe a person’s ability to understand the mental state of themselves and others [32]. It has been shown that depressed patients have deficits in their theory of mind [23][44]. It has also been shown that people with bipolar disorder (type I and II) have deficits in both theory of mind and facial expression recognition [28]. Additionally patients with borderline personality disorder (emotionally unstable personality disorder) have trouble correctly recognizing and expressing emotions [8], leading to severe impairments in self and interpersonal functioning [2].

Research has shown that facial expression training can have a positive effect on reducing irritability, negative interpretation bias and symptoms of depression and anxiety [40][26]. A major part of effective treatment for borderline patients (Dialectical Behavior Therapy) [27] consists of learning to correctly recognize and communicate emotions [35].

Using games to increase motivation
In psychology two types of motivation are defined: intrinsic and extrinsic motivation, a definition of these types of motivation has been described by Ryan and Deci [36]. Intrinsic motivation motivates people to do something because it is enjoyable. Extrinsic motivation sets people to do something because of the desired outcome. Traditionally the educational system refers to extrinsic motivation; spending time in school is not typically enjoyable, but the long term outcome is. Games on the other hand are more intrinsic in their way of motivating; they are enjoyable but have limited long term benefits. Przybylski et al. [34] showed how the psychological model for intrinsic motivation can be applied to games. Specifically, the three fundamental human needs that drive intrinsic motivation: the need for competence, autonomy and relatedness. The need for competence is often satisfied in games by slowly increasing difficulty and positive feedback. The need for autonomy is satisfied by giving players a freedom of choice. And finally the need for relatedness is often satisfied by having social interaction in the game usually by allowing people to play together.

Games have been analyzed in a teaching context and shown to increase children’s motivation, while still reaching educational goals [33][24]. Additionally games can also reduce the focus on grades [42], possibly reducing performance anxiety. However, there has been little research on the long term benefits in real life (distant generalization) for educational games [18].

Games for training emotion recognition
Games rarely contain real actors; the dynamic nature of games requires the characters to be adaptable. Evidence suggests that people show similar behavioral and physiological reaction to facial expressions in human-like virtual characters and recordings or photographs of real humans [11]. It was shown by Dyck et al.[13] that the expression of basic emotions in virtual faces can be recognized as well as in natural faces. This means that virtual characters in games are not a limitation to train facial expression recognition.

Because of their troubles with recognizing and using facial expressions, many applications and games have been aimed at people with ASD. While many studies show encouraging results in using computer programs to train emotions, there is little significant research to show that the improvements extend beyond the context of the programs or how it relates to clinically accepted therapies [30][21]. Most emotion training
applications have only been tested on children with ASD, but it could also be beneficial for neurotypical children.

The Transporters
One of the more promising research projects in the area of emotion recognition for children with ASD is The Transporters [19]. The Transporters is an animated series to train emotion recognition for children with ASD. In the series recorded faces of actors have been superimposed over animated vehicles. An example can be seen in figure 1. The DVD contains fifteen episodes of five minutes, each focusing on another emotion. Every episode is accompanied by a quiz to allow the children to practice what they have learned. The study compared a group of children with ASD (n=20) that used the DVD for four weeks (minimal three episodes per day) with two control groups: one with ASD (n=19) and one neurotypical (n=18). All participants were approximately 6 years old.

The study showed that the group that used the DVD, while performing worse than the neurotypical control group before the intervention, improved to the same level after four weeks (η² = 0.45 and 0.56). The group did not only improve in tasks where characters from the DVD were used (close generalization), but also with new situations and faces (distant generalization).

The authors claim that successful distant generalization was particularly interesting since earlier studies [7][38] had limited success in achieving this. However, the authors note that even though these results seem promising, they have not investigated if this distant generalization extends to the social functioning of the participants in real life situations.

A follow up study [43] replicated the experiment and conducted 1- and 3-month post intervention assessments. This study showed limited success in distant generalization; there was only an improvement in identifying anger immediately after the intervention, but not at follow-up. The writers do mention that the children assessed in this study were lower functioning (lower range of cognitive ability, more severe autistic symptoms) than those in the original trial. According to the study there is a positive correlation between age and verbal skills and the improvements achieved by the intervention, suggesting that the program is more effective for older, higher functioning children.

A recent study [17] with high functioning children with ASD, with one group with ASD that watched the DVD (n = 29) and both an ASD (n = 14) and a typically developing (TD) control group (n = 25) showed good long-term distant generalization results. Before the intervention the TD group performed significantly better than both ASD groups. After the intervention and at 3 month follow up the group that watched the DVD performed as well as the TD control group and significantly better than the ASD control group (η² = 0.22 and η² = 0.21). Additionally this study showed that the role of parental support was not significant, providing an argument for the use of the program in classrooms.

JeStimule
JeStimule is a prototype 3d game aimed to teach social communication to children and adolescents with ASD [37]. A picture from the game can be seen in figure 1. Players can freely walk around in a city where they can interact with objects and people. When a player interacts with a person a social scene is played, after the scene the player has to correctly identify the expressed emotion. Emotions were shown using facial expressions and/or gestures. Players were rewarded for correct answers with a piece of a puzzle. A unique element of this study is that the player is assessed on social communication abilities before the game and the method of answering is adapted to the results of the assessment. The methods included were: color codes for emotions (based on Plutchik’s wheel of emotions [31]), text labels or figurative descriptions of the emotions. Participants (n = 33) played the game eight times for maximum one hour over a period of four weeks. Emotion recognition in virtual characters, real-life characters and real-life characters in context was measured two weeks before and after training with the game. The participants showed an increase in their ability to recognize emotions in virtual characters (before: µ = 31.16, σ = 13.34, after: µ = 64.65, σ = 17.50, p < 0.001) and real-life characters (before: µ = 37.39, σ = 20.96, after: µ = 61.21, σ = 28.83, p < 0.001). The adaptability of the game to the cognitive ability of the players allowed both low functioning and high functioning children with ASD to be able play the game and get positive results from the intervention.

The JeStimule study shows that both high function and low functioning children with ASD can benefit from the program. However, as the authors state, having a non homogeneous group of subjects leads to greater variability in the results, and the specific differences are not explored in the paper. There was also no comparison with a control group and the effects on real life situations was not investigated.

Research questions and hypothesis
The main question this study sets out to answer is: what factors influence engagement and performance for children in a skill game to train facial expression recognition? It is hypothesized that performance will be influenced by facial expression intensity, context and the number and types of labels to choose from. It is expected that engagement will be influenced by the perceived difficulty and rewards for playing the game. Related to this, it is hypothesized that players’ performance influences engagement. An additional question is if there are specific expressions that are harder to identify correctly. It is hypothesized that happy and angry expressions are mostly identified correctly, surprise and afraid are often

Figure 1. Pictures from The Transporters (left) and JeStimule (right)
confused. And the final question this study sets out to answer is what children think about playing a game to train facial expressions. It is expected that children are positive about playing a game in a school setting, however they might expect the game to be more focused on enjoyment.

IMPLEMENTATION
The game was developed in the Unity1 game engine. Extra care was taken to make the game less graphically demanding, the program is supposed to run on older computers and mobile devices that are often found in schools. For this reason the game was developed in WebGL. This allows the game to run in a browser without requiring the user to install anything. Progress and performance are anonymously stored in a central database, making it possible for players to continue to play on a different computer. The central database also makes data collection easier.

Since facial expressions are dependent on context, several features were added to the game to add background to the faces. Expressions are shown in a story, a story consists of up to eight frames. Each frame shows one facial expression, a background image and a short paragraph of text. Frames have a certain duration (2, 5 or 10 seconds) in which the player has to select an answer. The time was limited to encourage players to go for their first impression, to make sure they would not change their answer after reasoning about it. The facial expression can have three different intensities (low, medium and high). The backgrounds consist of computer generated images of public spaces, such as city squares, restaurants and parks. There are also several pictures of classrooms and school hallways. Additionally, there are two modes the game can be played in: valence mode and naming mode. In valence mode players have to identify an emotion as being negative or positive. In naming mode four labels with basic emotions are shown and a user has to select the label that corresponds with the shown expression.

The basic gameplay loop is as follows: the frame is shown with the corresponding background image and text, the facial expression is neutral. When the player has read the text and is ready to continue they click on a button. A timer counts down from 3 and at 0 the facial expression changes into one of the six expressions. Now the player has limited time to select the right label. After selecting a label feedback is given and the next frame is shown.

A story editor is also provided to allow teachers or children to make their own stories. In this way the game can be used to reflect and learn from personal experiences.

Characters
For the facial expressions 3d animated characters were used, these can be easier manipulated and acquired than real faces from actors. The characters used in the game are made by Morph3D2. The packages used in the game are MCS Male and MCS Female in combination with the freely available outfit bundles MCS Female Sci-Fi Bundle and MCS Male: Wood Elf Scout.

The three free outfit bundles and the original characters contained eight outfits in total (six for the female and two for the male character). Of the female outfits only three were deemed appropriate as the rest showed too much skin. Additionally, all the weapons were removed from the characters to fit the educational context. The outfits that are in the game can be seen in figure 2.

The MCS characters allow developers to change the facial expressions using morphs. These morphs refer to Unity’s internal blendshapes; values that change the shape of the mesh. They can be set either before or during runtime and change the shape of the mesh dynamically. The morphs were used to change the facial expressions at runtime, using predefined values from a database. The library with values for the facial expressions was available from an earlier project and have not been properly validated. The expression that were added to the game are the six basic emotions: anger, disgust, fear, happiness, sadness and surprise. The expressions can be seen in figure 3.

Performance feedback
There are two ways the user gets feedback on his or her performance. The first is directly after selecting an answer in the form of stars that appear, the second is a score overview after playing through a story.

Every time a player answers a question a visual effect is displayed: several golden stars jump from the selected answer and fall. The number of stars that appear is larger when more people have selected the same answer. This way the user gets direct feedback on his or her performance. Figure 4 shows a player selecting a correct answer.

Since the labeling of facial expressions might be culturally influenced, no default right or wrong answers have been set. Instead, the program saves the user’s score in a central database and the score is calculated based on the percentage of people that gave the same answer as the user. For the testing of the program, the database was filled with percentages, with a higher percentage for the intended emotion and lower percentages for the other emotions. For a potential future release of the game the database should be filled with either a trial group, or facial expression experts.

Rewards
The players collect points while playing the game. The point system has been designed to encourage the players to keep playing, regardless of their performance. As stated before, the score is always a percentage based on what other users selected. Secondly, users get points for the number of frames in a completed story. So even in the hypothetical situation that a player gets no points from the selected answers, he or she will still receive some points.

Outfits were added to the game and can be unlocked using the points that the users acquire. This adds another incentive to keep playing to the game. Players can choose to unlock a new outfit for either the male or female character. Each

---

1https://unity3d.com/
2https://www.morph3d.com/
METHOD
The evaluation of the game was performed with typically developing children on a public school in a city in Sweden. The trial was performed on two dates, one week apart. The children were asked to fill in a questionnaire after the first trial. There was a bug in the software during the first trial that caused the facial expressions to not appear correctly. This was resolved for the second trial a week later. The second trial also included several other changes, most notably a Swedish translation and a forced order to play the stories. The children played the game for approximately 60 minutes in each trial.

Participants
54 students (8 years old; \(n = 31\), 11 years old; \(n = 23\)) participated in the study (See table 1). They played the game in groups of two or three. There were 29 groups of children playing the game, divided by age. Every group received an made up name to identify them with. One identifying name was used in both age groups, so it was not clear to which group the results belonged, this group was excluded from the results. Resulting in a total of 27 groups, 15 groups of 8 year old students and 14 groups of 11 year old students.

Stimuli
The participants were presented with 2 groups of 6 stories that had to be completed in order. The first 6 stories consisted of 5 frames where participants were asked about the valence of the emotion, this was called “Valence-mode”. The expression surprised was excluded from this because it has a neutral valence. In the second 6 stories of 6 frames the participants were asked to pick the shown emotion from 4 labels, this was called "Naming-mode". Within each group of stories

<table>
<thead>
<tr>
<th>Age</th>
<th>Female</th>
<th>Male</th>
<th>Unknown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 years old</td>
<td>11</td>
<td>15</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>11 years old</td>
<td>8</td>
<td>13</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>19</strong></td>
<td><strong>28</strong></td>
<td><strong>7</strong></td>
<td><strong>54</strong></td>
</tr>
</tbody>
</table>
the intensity was decreasing, starting with 2 stories on high intensity, then 2 on medium intensity and finally 2 on low intensity. The stories were also alternating between congruous context and ambiguous context. With congruous context the text showed a line of text that related to the shown emotion, with an ambiguous context the text was unrelated to the shown emotion. The distribution of the modes and stimuli can be seen in Appendix A: Stimuli.

RESULTS

Engagement

The mean number of completed stories was 8.48 ($\sigma = 2.98$) from 12 stories that were provided. An overview of the number of completed stories can be seen in figure 5. 85% of the groups completed all the valence stories and 30% of all groups completed all stories. An independent-samples t-test showed no significant differences for the mean number of completed stories between age groups. Due to a bug data on the unlocked outfits was not stored properly and could not be analyzed.

Performance effect on engagement

Since all the groups completed the first three stories, a correlation test (Spearman) was performed to see if there was any relation between the number of correct answers in the first three stories and the number of completed stories. A medium significant correlation was observed ($r = 0.429$, $p = 0.025$). Looking at the groups that did not start naming (n=6), there was a non-significant negative correlation, $r = -0.739$, $p = 0.094$. For the groups that did finish valence (n=21) there was a strong, significant positive correlation, $r = 0.659$, $p = 0.001$. However, for the groups that did not finish the whole game (n=13) this correlation is no longer significant, $r = 0.356$, $p = 0.233$. A Mann-Whitney U test shows that the groups that finished all the stories did significantly better, during the first three stories, than the rest of the groups, $U = 29$, $Z = -2.616$, $p = 0.011$. 80% of these groups had a full score in the first three valence stories, of the groups that did not start naming 33% had a full score and of those who did start naming and did not finish the game this was 23%.

When looking at performance in the first naming story, for the groups that did finish valence (n=21), there is no significant correlation between performance and number of completed stories, $r = 0.213$, $p = 0.355$. However, when only looking at the groups that did not finish all the stories (n=13), there is a significant positive correlation between the performance in the first naming story and the number of completed stories ($r = 0.711$, $p = 0.006$). There was no significant difference for the score in the first naming story between these two groups.

Performance

Of the 29 groups that participated in the study, 23 completed all the stories in the valence-mode. No significant difference was found between age groups for the overall performance $[r(21) = -0.259, p = 0.798]$ and there was no significantly better performance for specific emotions (within groups or between groups). In table 2 the mean number of correct answers in the valence-mode for these groups can be seen.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Number of questions</th>
<th>8 years old (n=12)</th>
<th>11 years old (n=11)</th>
<th>All (n=23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High intensity 10</td>
<td>8.17 (0.94)</td>
<td>8.09 (0.94)</td>
<td>8.13 (0.92)</td>
<td></td>
</tr>
<tr>
<td>Medium intensity 10</td>
<td>9.33 (0.66)</td>
<td>9.64 (0.67)</td>
<td>9.48 (0.67)</td>
<td></td>
</tr>
<tr>
<td>Low intensity 10</td>
<td>8.25 (1.06)</td>
<td>8.36 (1.03)</td>
<td>8.30 (1.02)</td>
<td></td>
</tr>
<tr>
<td>Congruent 15</td>
<td>14.25 (0.87)</td>
<td>14.55 (0.93)</td>
<td>14.39 (0.89)</td>
<td></td>
</tr>
<tr>
<td>Ambiguous 15</td>
<td>12.83 (1.75)</td>
<td>12.82 (2.64)</td>
<td>12.83 (2.17)</td>
<td></td>
</tr>
<tr>
<td>Collapsed 30</td>
<td>27.08 (2.07)</td>
<td>27.36 (3.07)</td>
<td>27.22 (2.54)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Mean number of correct answers in valence mode, split on intensity and text-mode. Standard deviation is between parenthesis.
To compare the mean score between congruous text and ambiguous text as paired-samples t-test was performed. The students scored significantly higher in congruous text mode \( t(22) = 3.527, p = 0.002 \). To see if certain emotions had a bigger influence on this difference than others, a paired-samples t-test was performed on every emotion, grouped by congruous text and ambiguous text. A significant difference was found for “Happy” (congruous \( \mu = 2.83, \sigma = 0.39 \), ambiguous \( \mu = 2.22, \sigma = 0.90 \), \( t(22) = 3.275, p = 0.003 \)). And “Disgusted” (congruous \( \mu = 2.87, \sigma = 0.34 \), ambiguous \( \mu = 2.52, \sigma = 0.67 \), \( t(22) = 2.336, p = 0.029 \)). The distribution of correct answers per emotion, per text-mode, can be seen in figure 6.

A paired-samples t-test between intensities showed that the groups scored significantly higher in medium intensity than both high \( t(22) = -6.916, p = 0.000 \) and low \( t(22) = 6.010, p = 0.000 \) intensities. There was no significant difference between high and low intensity \( t(22) = -0.848, p = 0.406 \). The distribution per emotion, per intensity, can be seen in figure 7. Paired t-tests were performed to determine significant differences between intensities within emotions. Significant differences were observed regarding “Happy” Medium and Low \( t(22) = 3.761, p = 0.001 \) as well as High and Low \( t(22) = 2.152, p = 0.043 \), “Sad” Medium and Low \( t(22) = 2.787, p = 0.011 \) and “Disgusted” High and Low \( t(22) = -2.299, p = 0.031 \). There were no significant age effects for these emotions.

For the groups that completed all the stories \( n = 8, 8 \) year old: \( n = 6, 11 \) year old \( n = 2 \), a paired-samples t-test was performed to compare the sum of correct answers in valence mode \( \mu = 28.38, \sigma = 1.41 \), to the sum of correct answers in the naming mode \( \mu = 16.13, \sigma = 3.18 \). The emotion “Surprised” was excluded from this comparison since it was not used in the valence part. The groups scored significantly higher in valence than in naming mode \( t(7) = 8.457, p = 0.000 \). In figure 8 the individual emotions are compared between valence and naming modes. Significant differences were found for Disgusted \( t(7) = 9.734, p = 0.000 \), Sad \( t(7) = 6.481, p = 0.000 \), Afraid \( t(7) = 9.721, p = 0.000 \) and Angry \( t(7) = 2.826, p = 0.026 \). This group did not have a significant difference between the sum of correct answers of connected text in valence mode and the sum of correct answers of ambiguous text in valence mode.

There was however a significant difference for the sum of correct answers in naming mode, with a higher score with congruous text \( \mu = 10.88, \sigma = 2.03 \) than ambiguous text \( \mu = 8.75, \sigma = 2.12, t(7) = 2.959, p = 0.021 \). For specific emotions “Afraid” scored significantly higher in congruous text mode \( \mu = 1.38, \sigma = 1.19 \) than in ambiguous text mode \( \mu = 0.12, \sigma = 0.35, t(7) = 3.035, p = 0.019 \). However, “Happy” scored significantly lower in congruous text mode \( \mu = 2.50, \sigma = 0.53 \) than in ambiguous text mode \( \mu = 2.88, \sigma = 0.35, t(7) = -2.049, p = 0.080 \).

Within the naming part groups scored significantly lower with low intensity \( \mu = 5.50, \sigma = 2.14 \) than medium \( \mu = 7.00, \sigma = 0.93 \), \( t(7) = 2.646, p = 0.033 \) and high \( \mu = 7.13, \sigma = 1.25, t(7) = 2.489, p = 0.042 \) intensities.

No age comparison could be made because of the low number of groups that made it this far.

A confusion matrix was made to show misattribution per shown emotion, collapsed across intensities and text-mode, for
the groups that completed part of the naming stories (n=21). The matrix can be seen in figure 9. On average 57% of the 8 year old and 63% of the 11 year old groups identified the expressions correctly.

**Questionnaire**

The questionnaire was presented after the first trial with the game. Note that the performance results presented before were from the second trial where several bugs were fixed and changes were made. Of the 54 students (8 year old; n = 31, 11 year old; n = 23) that participated in the questionnaire 7 did not fill in their gender and were excluded from the analysis (8 years old; n = 5, 11 years old; n = 2). Analysis was performed on the remaining 47 students.

**Description of the game**

The participants (n = 47) were asked to associate words with their experience of playing the game. There were 5 positive words ("nice", "easy", "exciting", "very good" and "fun"), 6 negative words ("childish", "difficult", "confusing", "boring", "ugly" and "bad") and 1 neutral word ("surprising"). They could select as many words as they wanted. The children selected more positive words about the game (µ = 1.57, σ = 1.43) than negative (µ = 0.84, σ = 0.84). t(46) = 2.837, p = 0.007. There was no significant difference between age or gender groups.

**Rating**

In this part of the questionnaire the students had to mark to what extent they agreed with the posed questions on a scale from 1 to 5.

Of the 47 remaining students another 10 students did not answer one or more questions for this part and were also excluded (8 year old female: n = 2, 8 year old male: n = 3, 11 year old female: n = 2, 11 year old male: n = 3). The questionnaire results for the remaining users that answered all questions (n = 37) can be seen in table 3.

An independent-samples t-test was conducted to compare the questionnaire results between age groups (8 years old: n = 21, 11 year old: n = 16). There was no significant difference for the overall rating in the questionnaire, but further investigation showed significant differences between specific questions. A significant difference was found for the players that reported that they finished the game where 8 year old students rated this lower than 11 year old students, t(30) = −2.495, p = 0.018. 8 year old students also reported feeling less in control while playing the game than 11 year old students, t(30) = −2.663, p = 0.012.

An independent-samples t-test between gender groups did not show any significant results.

However, an independent t-test between gender groups within age groups showed a significant difference for the 11-year-old students (female: n = 6, male: n = 10). 11 year old female students reported feeling more comfortable while playing the game than 11 year old male students, t(14) = −2.483, p = 0.026.

Another test for age groups within gender groups showed one significant difference for female students (8 year old: n = 9, 11 year old: n = 6) where 11 year old female students felt more comfortable while playing the game than 8 year old female students, t(8) = −2.345, p = 0.047.

**Open questions**

The questionnaire contained a section with three open questions where the students were asked what they liked about the game, what they did not like and what they thought the goal was to learn something in general.

**DISCUSSION**

The purpose of this study was to assess what factors affect engagement and performance in a social game to train facial expression recognition. The game was tested on a school in Sweden with 54 typically developing children, ages 8 and 11. Their scores and number of completed stories were recorded. Additionally the children filled in a questionnaire about their experience with the game. 3

3In the second trial a Swedish translation was provided.
<table>
<thead>
<tr>
<th></th>
<th>8 years old</th>
<th></th>
<th>11 years old</th>
<th></th>
<th>Overall</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Overall</td>
<td>Female</td>
<td>Male</td>
<td>Overall</td>
</tr>
<tr>
<td></td>
<td>(n=9)</td>
<td>(n=12)</td>
<td>(n=21)</td>
<td>(n=6)</td>
<td>(n=10)</td>
<td>(n=16)</td>
</tr>
<tr>
<td>Did you like the game?</td>
<td>4.00</td>
<td>3.75</td>
<td>3.86</td>
<td>3.83</td>
<td>4.40</td>
<td>4.19</td>
</tr>
<tr>
<td></td>
<td>(1.00)</td>
<td>(1.29)</td>
<td>(1.15)</td>
<td>(1.17)</td>
<td>(0.52)</td>
<td>(0.83)</td>
</tr>
<tr>
<td>Did you finish the game?</td>
<td>4.22</td>
<td>3.67</td>
<td>3.90</td>
<td>4.83</td>
<td>4.60</td>
<td>4.69</td>
</tr>
<tr>
<td></td>
<td>(0.97)</td>
<td>(1.44)</td>
<td>(1.26)</td>
<td>(0.41)</td>
<td>(0.70)</td>
<td>(0.60)</td>
</tr>
<tr>
<td>Was the game easy for you?</td>
<td>3.44</td>
<td>3.08</td>
<td>3.24</td>
<td>4.00</td>
<td>3.50</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>(1.67)</td>
<td>(1.44)</td>
<td>(1.51)</td>
<td>(1.10)</td>
<td>(0.71)</td>
<td>(0.87)</td>
</tr>
<tr>
<td>Would you like to play the game again?</td>
<td>4.11</td>
<td>3.92</td>
<td>4.00</td>
<td>3.83</td>
<td>4.30</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>(1.05)</td>
<td>(1.31)</td>
<td>(1.18)</td>
<td>(1.17)</td>
<td>(0.67)</td>
<td>(0.89)</td>
</tr>
<tr>
<td>Did it feel like you could control the game?</td>
<td>3.56</td>
<td>3.25</td>
<td>3.38</td>
<td>4.67</td>
<td>4.20</td>
<td>4.38</td>
</tr>
<tr>
<td></td>
<td>(1.51)</td>
<td>(1.54)</td>
<td>(1.50)</td>
<td>(0.52)</td>
<td>(0.79)</td>
<td>(0.72)</td>
</tr>
<tr>
<td>Did the game respond quickly when you played?</td>
<td>3.33</td>
<td>3.50</td>
<td>3.43</td>
<td>4.50</td>
<td>4.10</td>
<td>4.25</td>
</tr>
<tr>
<td></td>
<td>(1.41)</td>
<td>(1.73)</td>
<td>(1.57)</td>
<td>(1.22)</td>
<td>(0.86)</td>
<td>(1.00)</td>
</tr>
<tr>
<td>Were the computer’s responses clear?</td>
<td>4.22</td>
<td>3.92</td>
<td>4.05</td>
<td>4.00</td>
<td>4.20</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>(1.39)</td>
<td>(1.44)</td>
<td>(1.40)</td>
<td>(1.26)</td>
<td>(0.79)</td>
<td>(0.96)</td>
</tr>
<tr>
<td>Did you feel like you were taking part in the game?</td>
<td>4.00</td>
<td>4.00</td>
<td>4.00</td>
<td>5.00</td>
<td>4.60</td>
<td>4.75</td>
</tr>
<tr>
<td></td>
<td>(1.32)</td>
<td>(1.28)</td>
<td>(1.26)</td>
<td>(0.00)</td>
<td>(0.97)</td>
<td>(0.77)</td>
</tr>
<tr>
<td>How fast did you get used to the game?</td>
<td>3.57</td>
<td>3.75</td>
<td>3.67</td>
<td>4.17</td>
<td>4.20</td>
<td>4.19</td>
</tr>
<tr>
<td></td>
<td>(1.42)</td>
<td>(1.54)</td>
<td>(1.46)</td>
<td>(1.17)</td>
<td>(1.14)</td>
<td>(1.11)</td>
</tr>
<tr>
<td>Did you feel comfortable while playing?</td>
<td>3.78</td>
<td>3.67</td>
<td>3.71</td>
<td>5.00</td>
<td>4.10</td>
<td>4.44</td>
</tr>
<tr>
<td></td>
<td>(1.56)</td>
<td>(1.78)</td>
<td>(1.65)</td>
<td>(0.00)</td>
<td>(0.88)</td>
<td>(0.81)</td>
</tr>
<tr>
<td>Did you like to play in a group?</td>
<td>3.89</td>
<td>4.08</td>
<td>4.00</td>
<td>4.50</td>
<td>4.30</td>
<td>4.38</td>
</tr>
<tr>
<td></td>
<td>(1.76)</td>
<td>(1.56)</td>
<td>(1.61)</td>
<td>(0.84)</td>
<td>(1.16)</td>
<td>(1.02)</td>
</tr>
<tr>
<td>Overall</td>
<td>3.83</td>
<td>3.69</td>
<td>3.75</td>
<td>4.39</td>
<td>4.23</td>
<td>4.29</td>
</tr>
<tr>
<td></td>
<td>(0.85)</td>
<td>(1.33)</td>
<td>(1.13)</td>
<td>(0.46)</td>
<td>(0.51)</td>
<td>(0.48)</td>
</tr>
</tbody>
</table>

Table 3. Mean results of the children’s rating of the game on a scale from 1 to 5, the standard deviation is between parenthesis.
Engagement
The children were generally positive about the game, they selected more positive words to describe the game and reported that they liked to play the game. The children also reported in the open questions that they liked to learn by playing and thought that the outfits were important. No age differences were observed in the number of completed stories in the game. But 11 year old students reported in the questionnaire that they felt more in control and performed better than the 8 year olds. Particularly 11 year old female students reported feeling more comfortable while playing.

Unfortunately it was not possible to analyze whether the outfits affected engagement. However, it was mentioned by a large amount of students both in the questionnaire and verbally that they liked the outfits and would like to see more which indicates that the outfits could be a motivational factor. It was also mentioned by players that they would like more interaction in the game, they compared it to commercial games like Grand Theft Auto and Mass Effect. While these games are not suitable for an educational environment, certain aspects of these games could be adapted to the fit this game. A game world where players are free to move around in and choose which characters they interact with, much like JeStimule [37]. This would satisfy the players need of autonomy more as stated by Przybylski et al. [34]. Some of these features could be combined with the label picking system of this game. However, the hardware restrictions of the computers at schools and making it possible to run the game in a browser might make this challenging.

The definition and measurement of engagement is complex [5], in this study the total number of completed stories was taken as a measurement of engagement. It was hypothesized that the players’ performance influences their engagement, lower achieving children might stop playing sooner than those who achieve higher scores. The results show that this was not always the case, there was a medium positive correlation between score and the number of completed stories. However, this seems to be mostly caused by the group of players who finished the game, who got a significantly higher score in the beginning of the game. The groups that did not finish the game have no significant correlation between score in the beginning of the game and number of stories played. Three groups of players were identified: those that quit before the valence mode was over ("Early Quitters"), players that played longer but did not finish ("Middle Group") and those who played until there were no stories left ("Completers"). These groups might have different factors that influenced their engagement and are analyzed separately.

Early Quitters
The first group, those who quit before the naming part did not get a lower score than the others, and their performance shows a negative trend with the number of stories they played. From this can be concluded that they did not stop playing because they found the game hard. It could have been that they found the game too easy and got bored of it. The promise of more outfits to unlock after playing longer was not enough to keep this group interested.

From the results of this study it is not possible to determine what could have motivated these students to play longer. However, since the players in this group did get high scores on average it could be interesting to see if they would have played longer with a more challenging game mode. To motivate this group to keep playing longer a faster progression from valence to more difficult questions could be implemented. This could satisfy their need for competence as described by Przybylski et al. [34]

Middle Group
These groups’ engagement did not seem to be affected by their initial performance in the game, but there was a correlation between the score in the first naming story and the number of stories they played. There was a high drop-out rate at this part of the game; after the first two stories 43% of all the remaining groups, which means nearly 70% of this group, had stopped playing. Of the groups in the Middle Group 30% quit after the first naming story. A potential reason for the high dropout after the valence section is the sudden increase in difficulty. The players were used to a specific way of playing where they only had to choose between positive and negative labels. The change to having to choose between four specific emotion labels might have been a too fast increase in difficulty, resulting in a decrease in performance and therefore motivation since the need for competence was not satisfied.

To motivate this group to keep playing longer a more gradual increase in difficulty could be implemented. This can be achieved by having an additional mode where players have to choose the correct emotion from just two labels, possibly with opposing valence, instead of four.

Although it was not possible in this study to analyze specific data about the outfits, many of the groups in this group would have been able to unlock all the outfits in the game. In the open question part of the questionnaire, several of the students (22%) mentioned that there were not enough outfits to choose from. A large number of students mentioned this also verbally to the teachers and supervisor. If the outfits were a reason to keep playing, it could lead to decreased engagement when there were no more outfits to unlock. The combination of decreasing performance in the absence of desirable rewards might have led to these groups dropping out.

Completers
For the third group the number of outfits did not play a big role to keep playing, they acquired more than enough points and kept playing. They also did significantly better than the other groups in the beginning of the game. From the results it is also visible that 50% of the groups that finished the game replayed early stories after finishing the game to get more points. It was observed during the trial that some of the students had competitions between each other, who could get the most points.

This group does not need the game to change much to keep them engaged. However, it is still possible to learn from the way these players played the game. Several students mentioned verbally that they were interested in having a high score list in the game so they could compare their scores with friends.
This is something that should be considered with care; while it might encourage high achieving students, it could have the opposite effect on players that have more difficulty with the game. The points system could also be reworked so that players receive more points for difficult stories, to prevent them from playing the easy, early sequences over and over again. This group could also be challenged more by extending the game with an additional, more difficult mode where all six alternative emotions are presented on the labels.

**Performance**
To answer the question about which factors made the game difficult for the children, performance in the game’s different modes was analyzed.

The valence mode was added to the game to provide players with an easy mode to play to familiarize themselves with the game and characters. The results show that the groups performed better in valence mode than in naming mode; this applied to all expressions, with the exception of happy. It was expected that the difference in intensities would influence the performance of the players, with a high intensity being easier and a low intensity harder. However, in valence mode the students performed significantly better with medium intensity expressions. In naming mode the hypothesis holds as the performance decreases with the intensity. A possible reason for the small difference between the intensities in the valence mode is the overall high score in this mode. A clear difference in performance between congruous and ambiguous text mode was observed, in both valence and naming modes. This confirms that context has a large influence in recognizing emotions as was also shown by Feldman Barrett et al.[6].

There was no major difference in average misattribution between the two age groups, however the wrong labels they selected was different between the groups. Overall the groups that did play in the naming mode were good at correctly identifying angry and happy. Both age groups were mostly correct in identifying surprised, but confused it in 30% of the cases with afraid. Afraid, on the other hand, was mostly labeled incorrectly as sad. The majority of the groups identified sad correctly, but 40% of the 8 year old groups confused it with angry and 25% of the 11 year olds with afraid. The 11 year old groups were mostly correct in identifying disgusted, but 8 year old groups identified it incorrectly as angry.

The results of this study correspond with the findings of Castelli [9]; happy is most often identified correctly, followed by anger. Surprise is often confused with fear. However, other than in the research by Castelli, the confusion for surprised is not symmetrical, fear is more often confused with sadness. Also, the confusion of disgusted and angry does not match earlier research. It is possible that this difference is caused by the facial expressions used on the virtual characters. The expressions were not validated by a large group of adults or professionals, so it could be possible that the expressions did not show the emotion that was expected.

**Valence**
A valence mode was added to provide the players with an accessible way to learn to play the game. The results show that the students were in general good at picking the correct valence for the shown expressions. However, as discussed, there were some issues with engagement that might have been caused by the valence mode. Furthermore, there are some more fundamental issues with the use of valence for facial expressions and emotions that should be considered for a future version of the game.

The division between expressions with positive and negative valence is not equal. Of the five expressions that were used in the valence mode, four had negative valence and only happy has positive valence. Even when adding complex emotions to the game (shame, pride and guilt), the distribution between positive and negative emotions is not symmetrical. There are more words that describe negative emotions than positive [3]. Green [20] argues that the reason for the larger differentiation in negative emotion labels comes from the need to specify what causes a negative feeling to be able to change it, whereas with a positive feeling this is not required.

Additionally, the use of valence might be confusing to players. The valence was intended to reflect the feeling an emotion induces, a negative valence means that the emotion is experienced as uncomfortable, while a positive valence means that the emotion is experienced as pleasant. However, marking an expression as “positive” or “negative” might teach the children that it is considered bad to show certain emotions. For example, fear has a negative valence, but the primary purpose for fear is to warn for danger. As such fear should not be considered a negative emotion. The same applies to sadness, while having a negative valence, it is not good or bad to show sadness. A risk is that the children learn to associate the word “bad” with emotions with negative valence, and as a result start to suppress these feelings. Research shows that both positive and negative emotions play an important role in psychological well being [1]. A possible solution is to rename the labels in the valence mode, instead of using “good” and “bad”, the labels could show “feels good” and “feels bad” as used in a study by Kestenbaum [25].

**Future work**
More research is needed to assess the long term generalization effects of game based interventions to teach emotions. As Dormann et al. [12] state, many commercial games contain elements that can be used for socio-emotional learning, but do not use them for specifically that goal. It would be interesting to see a commercial game that is both enjoyable and has the socio-emotional learning as a main goal.

This game is currently under further development by the Embodied Social Agents Lab and it will be used in further studies. The game is planned to be used as a support tool in an classroom setting. Based on the observations of this study it was apparent that the game is usable in an one hour teaching session. However, more research is needed to develop a suitable lesson plan, for this it is particularly relevant to assess the effectiveness of the game in a classroom setting and the teachers’ acceptance of games in the educational program. It is also

4<http://www.csc.kth.se/~chpeters/ESAL/>
relevant to know what motivates the identified groups to keep playing and what could be changed to enhance engagement.

It is planned to expand the database in the game with statistics on how often a label was picked with a shown expression, these numbers will determine how many points a player gets for his answer. This database needs to be filled by a group of adults or experts on facial expressions. Emotions are affected by context [6], in the current version of the game the only context consists of the text and the background. Full body postures and complex emotions could be added to the game to include a broader context.

Limitations
The biggest limitation of this study is that the order of the stories and the ordering of the expressions and intensities within the stories was fixed. This was done to allow a progression from expected easy to difficult tasks. However, the fixed order could have led to ordering and learning effects that influence the results of this study. For example, children might have needed some time to get used to playing the game. Since the first frame was always High intensity Disgusted with congruent text in valence mode, the performance for this particular combination might have been lower. The same applies to combinations that were later in the game, where fatigue might have played a role. For a future study this could be avoided by using a random order of stories and frames within stories.

There were several additional limitations. The participants did the experiment in groups, this means that nothing can be said about individual students’ performance and engagement. The trial took place in a classroom setting, which is an uncontrolled environment. The questionnaire results were not linked to the participants of the experiment, this means that it is not possible to find a relationship between how the game was perceived and how the students performed. Furthermore, the questionnaire was only done at the first trial when the game still had some significant bugs. Only three other emotions were shown per emotion in the naming mode, this made it impossible to analyze all possible confusing emotions. Due to a restriction in the WebGL implementation of the characters it was not possible to animate the faces. There are also no whole body poses to accompany the facial expressions. Both of these points remove the game from natural social communication, making generalization harder. The facial expressions of the virtual characters were not validated with a large population, this means that is not known how realistic the expressions were compared to expressions used by real humans. Finally, the learning effect of the game was not tested for short or longterm distant generalization, meaning that nothing can be said about the effectiveness of the game to teach expression recognition for every day social communication.

CONCLUSION
The study shows that determining a facial expression to have a positive or negative valence is significantly easier than correctly identifying the shown expression. In fact, there are indications that the valence mode might have been too easy compared to the naming mode, seeing that students some students dropped out early despite a good performance. The intensity of the facial expressions influenced the performance, but only in naming mode. As expected, the children were better at identifying an expression in a matching context. The participants were generally good at correctly identifying happy and angry, this corresponds with earlier research. However, inconsistent with earlier research, the confusion in the other emotions was often not symmetrical. This might be caused by the expressions of the virtual characters and these should be validated in future research.

It did seem that the students that did exceptionally well in the beginning of the game played the whole game. However, performance later in the game did not affect this group as much as those who did not finish the game. Groups that quit early were probably not motivated because they did not find the game either enjoyable or challenging enough. Children that did finish the game most likely also had external motivation, such as competition with their peers. No conclusions could be drawn on the effectiveness of the outfits, but based on the feedback the participants gave it can be concluded that they appreciated the ability to customize their character.

The results of this study provide valuable information to further develop the game. Several of the proposed changes are: renaming the labels in valence-mode, having a faster and more gradual increase in difficulty with different amounts of labels, adding full body poses and adding more options to customize the character in the game. More research is needed to further analyze factors that influence engagement and the learning effectiveness of the game.

ACKNOWLEDGMENTS
I would like to thank my supervisor Christopher Peters for his help with brainstorming interesting research areas, providing me with the contacts needed for performing the experiment and guiding me through the project. Many thanks to Amanda Cederberg and the teachers at Holmaskolan in Malmö as well as the Embodied Social Agents Lab (ESAL) for conducting the experiment. I would also like to thank Robin Palmberg, Chengjie Li and Himangshu Saikia for providing me with an earlier version of the project, the facial expressions and helping me with programming and bug fixing. Thanks to the people at Scharc Arkitektur AB for providing me with a nice place to work away from the distractions of home. And finally, my biggest thanks go to my girlfriend Sofia Tingstam for her invaluable psychological and analytical insights, staying up with me all night to prepare for the user study, proofreading and her patience with me when I felt that there was no way I could put it all together.

REFERENCES


### Appendix A: Stimuli

<table>
<thead>
<tr>
<th>Story</th>
<th>Mode</th>
<th>Intensity</th>
<th>Context</th>
<th>Frame 1</th>
<th>Frame 2</th>
<th>Frame 3</th>
<th>Frame 4</th>
<th>Frame 5</th>
<th>Frame 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Valence</td>
<td>High</td>
<td>Congruous</td>
<td>Disgusted</td>
<td>Sad</td>
<td>Afraid</td>
<td>Angry</td>
<td>Happy</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Valence</td>
<td>High</td>
<td>Ambiguous</td>
<td>Afraid</td>
<td>Angry</td>
<td>Disgusted</td>
<td>Happy</td>
<td>Sad</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Valence</td>
<td>Medium</td>
<td>Congruous</td>
<td>Afraid</td>
<td>Happy</td>
<td>Angry</td>
<td>Sad</td>
<td>Disgusted</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Valence</td>
<td>Medium</td>
<td>Ambiguous</td>
<td>Angry</td>
<td>Disgusted</td>
<td>Afraid</td>
<td>Happy</td>
<td>Sad</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Valence</td>
<td>Low</td>
<td>Congruous</td>
<td>Angry</td>
<td>Disgusted</td>
<td>Sad</td>
<td>Happy</td>
<td>Afraid</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Valence</td>
<td>Low</td>
<td>Ambiguous</td>
<td>Sad</td>
<td>Happy</td>
<td>Angry</td>
<td>Afraid</td>
<td>Disgusted</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Naming</td>
<td>High</td>
<td>Congruous</td>
<td>Disgusted</td>
<td>Sad</td>
<td>Surprised</td>
<td>Angry</td>
<td>Happy</td>
<td>Afraid</td>
</tr>
<tr>
<td>8</td>
<td>Naming</td>
<td>High</td>
<td>Ambiguous</td>
<td>Afraid</td>
<td>Angry</td>
<td>Disgusted</td>
<td>Surprised</td>
<td>Sad</td>
<td>Happy</td>
</tr>
<tr>
<td>9</td>
<td>Naming</td>
<td>Medium</td>
<td>Congruous</td>
<td>Afraid</td>
<td>Surprised</td>
<td>Angry</td>
<td>Sad</td>
<td>Disgusted</td>
<td>Happy</td>
</tr>
<tr>
<td>10</td>
<td>Naming</td>
<td>Medium</td>
<td>Ambiguous</td>
<td>Angry</td>
<td>Disgusted</td>
<td>Afraid</td>
<td>Happy</td>
<td>Sad</td>
<td>Surprised</td>
</tr>
<tr>
<td>11</td>
<td>Naming</td>
<td>Low</td>
<td>Congruous</td>
<td>Surprised</td>
<td>Angry</td>
<td>Disgusted</td>
<td>Sad</td>
<td>Happy</td>
<td>Afraid</td>
</tr>
<tr>
<td>12</td>
<td>Naming</td>
<td>Low</td>
<td>Ambiguous</td>
<td>Sad</td>
<td>Happy</td>
<td>Angry</td>
<td>Surprised</td>
<td>Disgusted</td>
<td>Afraid</td>
</tr>
</tbody>
</table>

Table 4. Overview of the stories that were used during the trial.