



Huan, Rui (2021) *Evaluating child engagement in digital story stems using facial data*. PhD thesis.

<http://theses.gla.ac.uk/82040/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

Evaluating Child Engagement in Digital Story Stems using Facial Data

Rui Huan

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow



November 2020

Abstract

Engagement is a key factor in understanding people's psychology and behaviours and is an understudied topic in children. The area of focus in this thesis is child engagement in the story-stems used in child Attachment evaluations such as the Manchester Child Attachment Task (MCAST). Due to the high cost and time required for conducting Attachment assessments, automated assessments are being developed. These present story-stems in a cost-effective way on a laptop screen to digitalise the interaction between the child and the story, without disrupting the storytelling. However, providing such tests via computer relies on the child being engaged in the digital story-stem. If they are not engaged, then the tests will not be successful and the collected data will be of poor-quality, which will not allow for successful detection of Attachment status.

Therefore, the aim of this research is to investigate a range of aspects of child engagement to understand how to engage children in story-stems, and how to measure their engagement levels. This thesis focuses on measuring the levels of child engagement in digital story-stems and specifically on understanding the effect of multimedia digital story-stems on children's engagement levels to create a better and more engaging digital story-stem. Data sources used in this thesis include the observation of each child's facial behaviours and a questionnaire with Smiley-o-meter scale. Measurement tools are developed and validated through analyses of facial data from children when watching digital story-stems with different presentation and voice types.

Results showed that facial data analysis, using eye-tracking measures and facial action units (AUs) recognition, can be used to measure children's engagement levels in the context of viewing digital story-stems. Using eye-tracking measures, engaged children have longer fixation durations in both mean and sum of fixation durations, which reflect that a child was deeply engaged in the story-stems. Facial AU recognition had better performance in a binary classification for discriminating engaged or disengaged children than eye-tracking measurements. The most frequently occurring facial action units taken from the engaged classes show that children's facial action units indicated signs of fear, which suggest that

children felt anxiety and distress while watching the story-stems. These feeling of anxiety and distress show that children have a strong emotional engagement and can locate themselves in the story-stems, showing that they were strong engaged.

A further contribution in this thesis was to investigate the best way of creating an engaging story-stem. Results showed that an animated video narrated by a female expressive voice was most engaging. Compared to the live-action MCAST video, data showed that children were more engaged in the animated videos. Voice gender and voice expressiveness were two factors of the quality of storytelling voice that were evaluated and both affected children's engagement levels. The distribution of child engagement across different voice types was compared to find the best storytelling voice type for story-stem design. A female expressive voice had a better performance for displaying the 'distress' in the story-stem than other voice types and engaged children more in the story-stems. The quality of the storytelling voice used to narrate story-stems and animated videos both significantly affected children's levels of engagement. Such digital story-stems make children more engaged in the digital MCAST test.

Contents

Abstract	i
Acknowledgements	xiii
Declaration	xiv
Chapter 1 Introduction	1
1.1 Motivation.....	1
1.2 Research Questions.....	6
1.3 Thesis Statement.....	7
1.4 Thesis Structure.....	7
Chapter 2 Literature Review	9
2.1 Outline.....	9
2.2 The Definition of Engagement.....	9
2.2.1 Engagement in Human-Agent Interaction.....	10
2.2.2 Engagement in User-System Interaction.....	11
2.2.3 Engagement in Education.....	12
2.2.4 Child Engagement in Preschool Classrooms.....	14
2.2.5 Narrative Engagement.....	16
2.2.6 Discussion.....	18
2.3 The project focus – Manchester Child Attachment Story Task (MCAST).....	20
2.4 Multimedia Tools for Storytelling.....	25
2.4.1 Animation vs. Live-action video.....	26
2.4.2 Storytelling Voice.....	28
2.5 Measurement Methods.....	32
2.5.1 Self-report Measures.....	32
2.5.2 External Observation.....	37
2.5.3 Automated Measures.....	39
2.5.4 Automated Measure 1 – Eye-tracking Techniques.....	40
2.5.5 Automated Measure 2 – Facial Expression Recognition.....	42
2.6 Conclusion.....	44
Chapter 3 An Initial Study of Adult Engagement Measurements	46
3.1 Introduction.....	46
3.2 Methods.....	48
3.2.1 Participants.....	48
3.2.2 Procedure.....	48

3.2.3	Data Annotation	48
3.2.4	Data Selection	54
3.3	Eye-tracking measures	55
3.3.1	Fixation Identification	55
3.3.2	Fixation metrics.....	58
3.3.3	Classification.....	59
3.4	Results.....	60
3.4.1	Two Fixation Metrics	60
3.4.2	Classification.....	67
3.5	Discussion and Conclusions.....	68
Chapter 4	Child Engagement Measurements from Facial Data	72
4.1	Introduction	72
4.2	Methods.....	74
4.2.1	Participants	74
4.2.2	Procedure	74
4.2.3	Data Annotation	75
4.2.4	The Inter-Rater Reliability	77
4.2.5	Data Selection	80
4.3	Recognition of Child Engagement	81
4.3.1	Recognition using Eye-tracking.....	81
4.3.2	Recognition using Facial AUs	81
4.4	The Self-report Measure	82
4.5	Results of Eye-tracking Measures.....	83
4.5.1	Primary Fixation Metrics	84
4.5.2	Classification.....	90
4.6	Results of Facial AUs Recognition	91
4.6.1	Classification.....	92
4.6.2	The Most Frequent AUs of Engagement	96
4.7	Results of the Self-report Measure.....	97
4.8	Discussion	98
4.9	Conclusion	103
Chapter 5	Designing an Engaging Digital Story-stem	105
5.1	Introduction	105
5.2	Evaluating the effects of media type on engagement.....	107
5.2.1	Storytelling Voice	107
5.2.2	Live-action Videos vs. Animated Videos	111
5.3	Methods.....	113
5.3.1	Participants	113
5.3.2	Procedure	114
5.3.3	Data Annotation and Selection	114
5.4	Recognition of Child Engagement.....	114
5.5	The Self-report Measure	115
5.6	Results	116

5.6.1	Classification Performance	117
5.6.2	The Effect on Child Engagement of Storytelling Voices.....	121
5.6.3	The Effect of Child Engagement on Presentation Types	125
5.6.4	The Self-report Measure	127
5.7	Discussion	134
5.8	Conclusion	139
Chapter 6	Discussion and Conclusions	140
6.1	Introduction	140
6.2	RQ1: Can children’s spontaneous facial expressions be used to automatically measure engagement levels in digital story-stems?	140
6.3	RQ2: How do voice type and presentation type affect child engagement levels in digital story-stems?	144
6.4	Limitations and Future Work	148
6.5	Conclusion	151
Bibliography		154
Appendices		160
Appendix A	The User Engagement Scale (UES)	160
Appendix B	Items used for developing the narrative engagement scale [17]	161
Appendix C	Definitions of Levels and Types of the E-Qual Coding System	163
Appendix D	Information pack to children’s family	164
a)	Opt-out Consent Form	164
b)	Participant Information Sheet/ Letter to Child	165
c)	Letter to Parents/Carers	167
Appendix E	The smiley-o-meter questionnaire used for Chapter 4	169
Appendix F	The smiley-o-meter questionnaire used for Chapter 5	173
Appendix G	The instruction letter to labellers	175

List of Tables

Table 2-1. Definitions of engagement across different research areas.	19
Table 2-2. Six attributes of the User Engagement Scale (UES) [63].	34
Table 2-3. Four dimensions of measuring narrative engagement [17].	35
Table 2-4. Five subscales used for designing the questionnaire for children in this thesis.	36
Table 2-5. Various levels of engagement annotation scales from previous studies.	37
Table 2-6. Five considerations of a fixation analysis metric [71].	41
Table 3-1. (left) Overall engagement ratings (19 clips labelled as X due to quality). (right) The table of weights. The agreement across engagement levels 1, 2, 3, and 4: 84.16%: Cohen’s kappa = 0.825. Engaged data (level 3 and 4) agreement 88.95%: Cohen’s kappa = 0.748.	52
Table 3-2. The engagement level annotation categories used by the labellers.	53
Table 3-3. The final rating result of engagement level $l \in \{1, 2, 3, 4\}$ of 18 clips for participant P1 using the data selection procedure.	54
Table 3-4. The distribution of adult engagement levels, shown in the count of clips and in parenthesis in percent.	55
Table 3-5. The overall distribution of adult engagement and the total number of fixations, shown in counts and in parenthesis in percentages.	62
Table 3-6. Results of a Hochberg <i>post hoc</i> test to find the actual differences of the four engagement levels on the overall number of fixations per clip. *: The mean difference is significant at the 0.05 level.	63
Table 3-7. The overall distribution of the engagement levels and the total fixation duration, shown in counts and in parenthesis in percentages.	65
Table 3-8. Results of a Hochberg <i>post hoc</i> test to find the actual differences of the four engagement levels on fixation durations. *: The mean difference is significant at the 0.05 level.	66
Table 3-9. Description of accuracy metrics.	68
Table 3-10. Confusion matrix of the binary classifier of adult engagement (high vs low) using the fixation duration, shown in the number of classified clips.	68

Table 3-11. Accuracy metrics of the binary classification of adult engagement using fixation durations.	68
Table 4-1. The engagement level annotation categories used by the raters.	76
Table 4-2. Results of the agreement of annotating the clips in terms of child engagement using Fleiss kappa.	77
Table 4-3. Results of the agreement of annotating the frames in terms of child engagement using Fleiss kappa.	78
Table 4-4. (left) Overall Child engagement ratings (12 clips labelled as X due to quality). (right) The table of weights. The agreement across child engagement levels 1, 2, 3, and 4 agreement 84.19%: Cohen’s kappa = 0.793. Engaged data (level 3 and 4) agreement 91.87%: Cohen’s kappa = 0.802.	79
Table 4-5. (left) Overall child engagement ratings (4108 frames labelled as X due to quality). (right) The table of weights. The agreement across child engagement level 1, 2, 3, and 4 agreement 73.64%: Cohen’s kappa = 0.710. Engaged data (level 3 and 4) agreement 78.52%: Cohen’s kappa = 0.561.	80
Table 4-6. The distribution of clips in terms of child engagement levels, shown in the count of clips and in parenthesis in percent.	81
Table 4-7. The distribution of frames in terms of child engagement levels, shown in the count of clips and in parenthesis in percent.	81
Table 4-8. The items of the questionnaire and its related aspects.	83
Table 4-9. The overall distribution of child engagement and the total number of fixations, shown in counts and in parenthesis in percentages.	86
Table 4-10. Results of a Hochberg <i>post hoc</i> test to find the actual differences of the four child engagement levels on the overall number of fixations per clip. *: The mean difference is significant at the 0.05 level.	87
Table 4-11. The overall distribution of child engagement and the total fixation duration, shown counts and in parenthesis in per cents.	88
Table 4-12. Results of a Hochberg <i>post hoc</i> test to find the actual differences between the four child engagement levels on fixation durations. *: The mean difference is significant at the 0.05 level.	89
Table 4-13. Confusion matrix of the binary classifier of child engagement (high vs low) using the fixation durations as shown in the number of classified clips.	91
Table 4-14. Accuracy metrics of the binary classification of child engagement using fixation durations.	91

Table 4-15. Confusion matrix of the binary classifier for child engagement using facial AUs.	94
Table 4-16. Confusion matrices of the binary classifier for child engagement (high vs low) using the facial AU-intensity (left) and using AU-presence (right).....	94
Table 4-17. Accuracy metrics of the three classifiers for child engagement (high vs low) using the facial AU, AU intensity and AU presence respectively.	94
Table 4-18. The five most frequent facial action units (AUs) intensity and presence from the classification result, as shown in the number of frames and in parenthesis in per cents.	96
Table 4-19. Descriptive analysis of children’s answers for the questionnaire according to the four aspects of child engagement measurements.	97
Table 5-1. The pitch value of each storytelling voice type (FF = female flat, FE = female expressive, MF = male flat, ME = male expressive).	108
Table 5-2. Result of a Hochberg <i>post hoc</i> test to find the actual differences of the four storytelling voice types using the pitch values. *: The mean difference is significant at the 0.05 level.	109
Table 5-3. The four sentences that contain a sudden moment from the MCAST story-stems to be used to check the differences between the flat and expressive recordings.	110
Table 5-4. The allocation of children to watch the MCAST stories with different media types.	113
Table 5-5. The distribution of frames in terms of child engagement levels, shown in the count of clips and in parenthesis in percent.	115
Table 5-6. The items of the questionnaire and its related aspects.....	116
Table 5-7. Confusion matrix for the 4-class classification using the AU classifier, as shown in the number of frames.	118
Table 5-8. Confusion matrix for the 4-class classification using the AU intensity classifier, as shown in the number of frames.	118
Table 5-9. Confusion matrix for the 4-class classification using the AU presence classifier, as shown in the number of frames.	119
Table 5-10. Accuracy metrics of the AU classifier for child engagement level $l \in \{1, 2, 3, 4\}$ using each of the three classification architectures. The avg. was the average value for the performances of each individual class.....	120
Table 5-11. Accuracy metrics of the AU intensity classifier for child engagement level $l \in \{1, 2, 3, 4\}$ using each of the three classification architectures. The avg. was the average value for the performances of each individual class.....	120

Table 5-12. Accuracy metrics of the AU presence classifier for child engagement level $l \in \{1, 2, 3, 4\}$ using each of the three classification architectures. The avg. was the average value for the performance of each individual class.121

Table 5-13. Observed number of frames and marginals for the rows and columns by child engagement level and storytelling voice type. The marginals for the rows and columns were calculated by adding the frequencies across the rows and down the columns. .122

Table 5-14. The result of Chi-Square Test. ^a: 0 cells had expected count less than 5. The minimum expected count is 617.45 (shown in the crosstabulation, Table 5-15).....123

Table 5-15. The crosstabulation table between child engagement level and storytelling voice type (FE = female expressive, FF = female flat, ME = male expressive, MF = male flat). * Each subscript letter denotes a subset of storytelling voice type categories whose column proportions do not differ significantly from each other at the .05 level.123

Table 5-16. Observed number of frames and marginals for the rows and columns by child engagement level and presentation type. The marginals for the rows and columns were calculated by adding the frequencies across the rows and down the columns. .125

Table 5-17. The result of Chi-Square Test. ^a: 0 cells (0.0%) have expected count less than 5. The minimum expected count is 1240.6 (shown in the crosstabulation, Table 5-18)126

Table 5-18. The crosstabulation table between child engagement level and presentation type. * Each subscript letter demotes a subset of presentation type categories whose column proportions do not differ significantly from each other at the .05 level.126

Table 5-19. Descriptive analysis of children’s answers for the questionnaire according to the five aspects of child engagement measurements.....128

Table 5-20. Descriptive analysis for children’s answers for each aspect under the different storytelling voice types, shown with mean values and in parenthesis in std. Dev.....129

Table 5-21. Descriptive analysis for children’s answers for each aspect under the different video formats, shown with the mean values and in parenthesis in std. Dev.132

List of Figures

Figure 1-1. A <i>Manchester Child Attachment Story Task</i> (MCAST) [32] setup. The story-stem vignettes take place around a dolls-house with two dolls: one representing the caregiver (the left one) and the other the child (the right one).	2
Figure 1-2. The School Attachment Monitor (SAM) setup for administering the assessment.	4
Figure 2-1. An example of the mental model approach. The girl builds an image for a mental model that she has a nightmare when she was listening to a nightmare story. 17	
Figure 2-2. CMCAST computer interface [55]. Story-stems are represented by animations on the computer. Children are asked to watch the vignettes then complete each story by speaking to the computer and moving the dolls presented on screen using the mouse.	23
Figure 2-3. Two types of climax in a story [84]. The upper one shows a sudden climax while the bottom one contains an increasing climax.	30
Figure 2-4. The Smiley-o-meter based on a 5-point scale [73].....	37
Figure 2-5. Sample facial action units from the FACS (Full AUs see).	43
Figure 3-1. The way of splitting the video recording of one participant.	49
Figure 3-2. The rating form used for human labellers to record the engagement rating and engaged/disengaged behaviours.....	50
Figure 3-3. The rating scores of participant P1’s engagement levels watching the four MCAST story-stems from two independent labellers.....	51
Figure 3-4. Annotation results for the participant P1 taken from rating forms. Clips (S1~S18) were randomly annotated to five labellers (ID: L1~L5). The blank areas mean that no engaged/disengaged behaviours were recorded by labellers.	53
Figure 3-5. The heatmap of one participant’s gaze data and a screenshot of the corresponding story-stem.	56
Figure 3-6. Calculation of gaze angle [66].....	57
Figure 3-7. The screenshots of collected and aggregated data. (left) Raw gaze data was collected by the eye tracker. (right) The raw gaze data were grouped into fixations and combined with the engagement ratings from human annotation.	61

Figure 3-8. The number of fixations per clip was computed from the aggregated data. (left) Same as Figure 3-7 (right) to show the duration of each fixation in one 10s clip. (right) The pair of data (engagement, the number of fixations) was used in the ANOVA test.....	61
Figure 3-9. The overall distribution of the engagement levels and the average number of fixations per clip. The line shows the overall mean number of fixations of all fixations.....	62
Figure 3-10. The duration of fixations was computed from the aggregated data. (left) Same as Figure 3-7 (right) to show the duration of each fixation per clip. (right) The pair of data (engagement, the duration of fixations) used in the ANOVA test.	64
Figure 3-11. The overall distribution of the engagement levels and the average fixation durations. The line shows the overall mean fixation duration of all fixations.	65
Figure 4-1. A screenshot of the ‘nightmare’ story-stem.	75
Figure 4-2. A screenshot of the ‘illness’ story-stem.	75
Figure 4-3. An example of human annotation using a frame.....	76
Figure 4-4. The screenshots of collected and aggregated data. (top) Raw gaze data was collected by the eye tracker. (bottom) The raw gaze data were grouped into fixations and combined with child engagement ratings from human annotation.	84
Figure 4-5. The number of fixations per clip was computed from the aggregated data. (left) Same as Figure 4-4 (right) to show the duration of each fixation in one 10s clip. (right) The pair of data (child engagement, the number of fixations) was used in the ANOVA test.....	85
Figure 4-6. The overall distribution of child engagement levels and the average number of fixations per clip. The line shows the overall mean number of fixations of all fixations.....	86
Figure 4-7. The duration of fixation was computed from the aggregated data. (left) Same as Figure 4-4 (right) to show the duration of each fixation per clip; (right) The pair of data (child engagement, the duration of each fixation) was used in the ANOVA test.	88
Figure 4-8. The overall distribution of child engagement levels and mean fixation durations. The line shows the overall mean fixation duration in all clips.....	89
Figure 4-9. The screenshots of aggregated data. A subset of facial Action Units (AUs) was recognised by intensity (AU_r) and presence (AU_c) and combined with ratings of child engagement levels from human annotation.....	92

Figure 4-10. ROC curve of the binary classification for child engagement (high vs low) using the facial AU.	95
Figure 4-11. ROC curve of the binary classification for child engagement (high vs low) using facial AU-intensity (left) and AU-presence (right).	95
Figure 5-1. The pitch contour of sudden climax for the MCAST story-stems narrated by a female voice (blue = flat voice, orange = expressive voices), from left to right: nightmare, hurt knee, illness, and shopping	110
Figure 5-2. The pitch contour of sudden climax for the MCAST story-stems narrated by a male voice (blue = flat voice, orange = expressive voices), from left to right: nightmare, hurt knee, illness, and shopping	110
Figure 5-3. A group of screenshots of the MCAST story-stem vignettes displaying as a live-action video. Left: the nightmare story-stem; Right: the illness story-stem.	111
Figure 5-4. A group of screenshots of the MCAST story-stem vignettes displaying as an animated video. (The displayed story in animation was nightmare, hopscotch, illness and shopping from the top left one to the bottom right one respectively.).	112
Figure 5-5. The screenshots of aggregated data. A subset of facial Action Units (AUs) was recognised by intensity (AU_r) and presence (AU_c) and combined with ratings of child engagement levels from human annotation.....	117

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Stephen Brewster. I have known Professor Brewster since I was a master student in University of Glasgow, and it was a great honor to be his student for the master project. Without his inspiring instruction, kindness, patience and sense of responsibility, I could not continue my path of academic research in HCI. I respect him not only as a mentor, but also as a friend in my life. Besides my supervisor, I would like to thank all staffs and colleagues who give me help for everything during my PhD.

I would also like to thank my parents for their support. Their precious encouragement, trust, care and love motivate me in my study.

Declaration

With the exception of chapters 1 and 2, which contain introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1 Introduction

1.1 Motivation

Recent studies have emphasised that engagement is a key factor in understanding a user's psychology and behaviours in various areas such as videogames [13], education [28, 30, 53, 80], communication [38, 55, 74] and entertainment [82]. For instance, researchers have been working on implementing a conversational agent that adapts conversations with a user according to the user's engagement level to improve naturalness in human-agent communications [40]. In education, researchers have explored automated recognition of student engagement which may help teachers evaluate the engagement levels of their students to adjust the learning process appropriately [56, 88]. Xie *et al.* have investigated child enjoyment and engagement while doing puzzles to understand the design implications of tangible user interfaces [91]. Engagement in children is an understudied topic. The aim of this thesis is to investigate a range of aspects of child engagement to help understand how to engage children and how measure their engagement levels to see if the engagement was successful.

Miller *et al.* suggest that there are two issues that need to be considered when designing engagement experiments [54]. The first is the definition of engagement based on the specific purposes of research. The second is how to choose the proper methods to measure it. In these previous studies, there is no a general definition of the term "engagement" and it is interpreted based on different contexts and user groups of research. For example, engagement in human-robot interaction (HRI) is commonly defined from Sidner *et al.* [81] as "the process by which two (or more) participants establish, maintain and end their perceived connection". In education, student engagement has three components including behavioural engagement (a person's willingness to participate in a task), emotional engagement (a person's emotional attitude towards tasks) and cognitive engagement (a person's focused attention as well as creative thinking) [29]. For Xie *et al.*'s child users [91], engagement has been operationalised as "the amount of time that children spend interacting with their environment in a developmentally and contextually appropriate manner" [2].

One important area of focus in this thesis is engagement in the story-stem approach used in child Attachment evaluations. The story-stem approach is a reliable and valid assessment method for investigating the important relationships in a child's life, and has made significant contributions to Attachment theory [14, 15, 32, 77]. Attachment is the natural tendency of children to seek and to maintain the physical proximity with their caregivers (typically the mother) [87]. It is one of most important aspects of young children's relationship functioning, which provides protection and nurtures physical and psychological wellbeing. In this method, an interviewer gives the beginning of a story then asks the child to complete it, often acting out the scene using dolls. One instance of the story-stem approach is the widely-used *Manchester Child Attachment Story Task* (MCAST) [32] (Figure 1-1). MCAST is a standard child psychiatry test which uses structured doll play and short story stems to assess the attachment status of children and their caregivers [32]. During the MCAST, an interviewer shows a story-stem vignette to the child using two dolls, one representing the child and the other the mummy, and a dolls-house and asks the child to act out what happens in the rest of the story with the symbolic dolls. The way the child completes the story and behaviour during the test provides the cues necessary to assess the child's Attachment status.



Figure 1-1. A *Manchester Child Attachment Story Task* (MCAST) [32] setup. The story-stem vignettes take place around a dolls-house with two dolls: one representing the caregiver (the left one) and the other the child (the right one).

Engagement is vital in the initial phase of the test, where a child is given the beginning of a story by an assessor using the dolls. In this phase, the assessor aims to bring children into a deep engagement with the mildly stressful story (e.g. the child wakes at night alone with a

nightmare) to bring out their mental representation of attachment to their caregiver [32]. Engagement in the story-stem is important as it means that children focus on attending to the play and materials, are not distracted by other things, and feel empathy with the dolls and characters in the story. This is measured by a trained assessor's observation of facial expressions, using the MCAST protocol [32]. If a child is not emotionally engaged by the predicament shown in each story-stem, the psychiatrist cannot assess their attachment status based on their story and behaviour during the activity.

Unfortunately, conducting MCAST assessments is expensive and time-consuming. Examiners must attend high-cost courses followed by lengthy reliability training to be certified for MCAST. Furthermore, the efficiency of MCAST assessment is limited by the number of children that they can reach. Trained assessors must spend time observing children's facial expressions from video recordings and rating the child's engagement levels, which takes a long time [49]. This means that few children are tested. Early diagnosis of attachment problems makes treating the condition more straightforward. If untreated, it can lead to many problems later in life, from aggressive behaviour to cardiovascular disease [38].

To reduce the time and cost required for MCAST administration and assessment, a system called the School Attachment Monitor (SAM) is being developed, which is designed to automate attachment assessments by administrating the MCAST assessment and automatically classify attachment patterns [78]. SAM is a computer-based tool that can potentially measure parent-child attachment across the population in a cost-effective way. One of the challenges of computerising doll-play based assessments, such as SAM, is to successfully digitalise the interaction between the child and the story without disrupting the storytelling.

In SAM, the story-stems are presented on a laptop screen. During the SAM test, the child is asked to watch a video where an actor performs an MCAST story-stem vignette, and asks the child to complete the story with the help of two dolls [78] (Figure 1-2). In this way, people without MCAST training, such as teachers, could administer the SAM test to reduce the cost and involvement of fully trained MCAST administrators so that the efficiency of Attachment assessment could be improved by increasing the number of children to be reached.

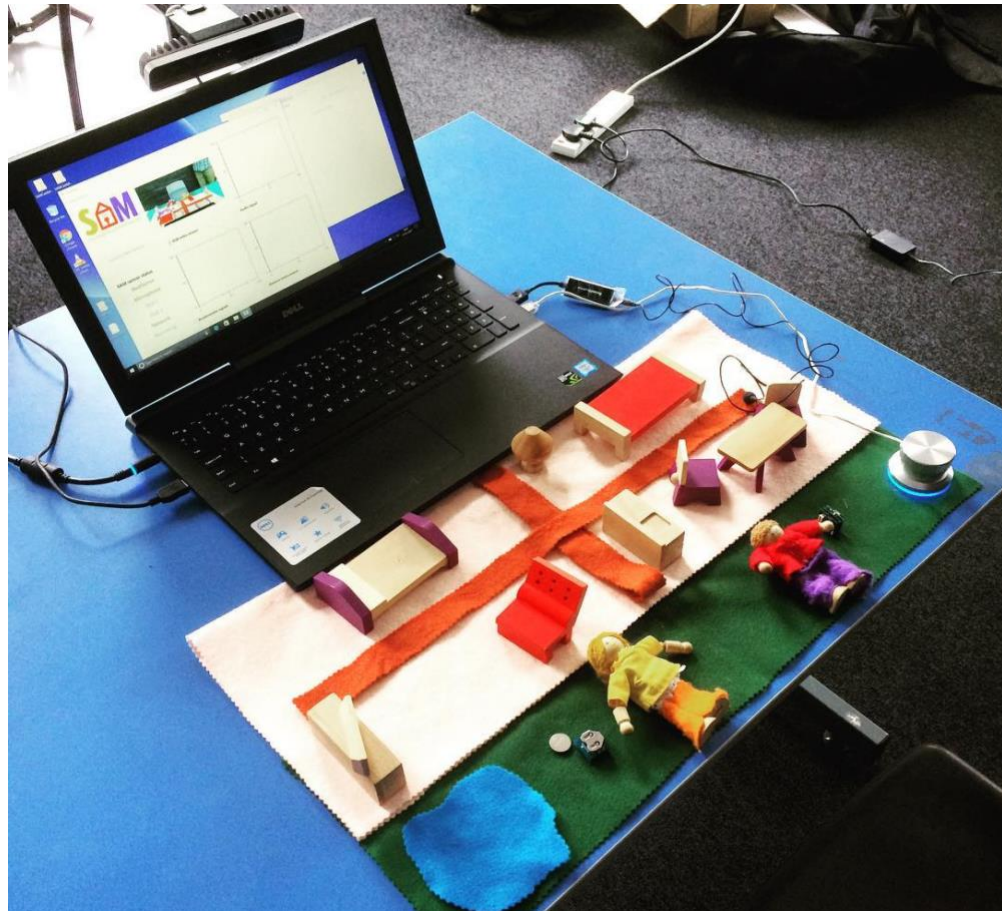


Figure 1-2. The School Attachment Monitor (SAM) setup for administering the assessment.

Engagement in SAM has the same important role as it in MCAST. An engaged child could complete the story to bring out their mental representation of attachment to their caregiver. If a child is not engaged in the video of a story-stem, SAM cannot analyse their attachment status based on their story and behaviour during the activity. Therefore, one aim of this thesis is to detect if children could be engaged in the digital story-stems used in the SAM study. As it is an experiment related to the term engagement, firstly the definition of engagement in the story-stem should be given here based on the MCAST protocol: *engagement in this thesis is a focusing of children's mood state around the particular distress represented in the MCAST story-stem*, as it means that children focus on attending to the play and materials, are not distracted by other things, and feel empathy with the dolls and characters in the story.

The second problem of engagement research taken from [54] is to choose the proper methods for engagement measurements that match the definition. There are three popular tools for measuring engagement: self-reports, external observation and automated measures. The self-report measure represents a robust and efficient approach by collecting users' perception using a questionnaire and/or interview to assess their engagement states by users' expressing their attitudes, feelings, beliefs or knowledge about a subject or situation [51, 56, 61, 89].

External observation is a common method for measuring child engagement [32, 43, 77]. Human observers are asked to follow checklists of measures that are supposed to indicate engagement. In MCAST assessment, child engagement is measured by a human assessor's observation of a child's facial expressions, using the MCAST protocol. One kind of automatic engagement recognition is based on computer vision, which provides an automatic estimation of engagement by analysing cues from the face and gestures [31, 50, 56, 88, 95]. In many situations, these kinds of behaviours are relatively easy to collect. For example, not looking at the TV can be a good indicator of low engagement while looking at it can be recognised as high engagement in a viewing task [36]. However, these methods were mainly designed for adults and there is little research [50, 95] related to the analysis of children's spontaneous facial expressions.

Therefore, the first stage of work in this thesis is a preliminary study for measuring adult engagement in MCAST story-stems to develop a scale for coding engaged behaviours and a method for capturing spontaneous facial expressions (e.g., eye movement and facial actions) that indicate engagement. These findings are then tested to measure children's engagement levels while watching the digital stories, which aims to answer RQ1:

Can children's spontaneous facial expressions be used to automatically measure engagement levels in digital story-stems?

Children are asked to watch the digital MCAST story-stems on a screen then to complete each story by playing with dolls and speaking to the computer with a web camera. This produces a video recording of children's facial expressions, which are used to measure the engagement level. Several face features are extracted and used to create a tool to identify children's engagement levels with MCAST stories. This method could be used for the SAM test to monitor the children's engagement levels successfully and identify 'disengaged' children. This reduces the time and effort of MCAST/ SAM coding and helps researchers know whether children are attending to the story and engaged in it; if they are not, then the test will not be successful and the collected data will be recognised as poor-quality, which will not allow for an accurate MCAST/SAM assessment.

Since engagement is an important concept in the tests using story-stems, bringing a child into a deep engagement while watching the digital story-stems vignettes could reduce the chance of poor-quality data assessment. Therefore, a demand here is focused on creating an

engaging MCAST digital story-stem vignette to make children more easily get absorbed to complete the vignette in spontaneous play. The use of multimedia technology, such as images, text, and recorded audio narration, could help children build their mental model for imagery of the story to improve their attention and story comprehension [17]. Different media types could be used to design MCAST stories including aspects of the storytelling voice such as gender and expressiveness, or aspects of the presentation such as animation, or live-action video. These were investigated in the thesis to answer RQ2:

How do voice type and presentation type affect child engagement levels in digital story-stems?

For example, audiences may not be consistently engaged where a narrator's voice tone does not fit the story line or a narrator's voice has a flat tone [86]. Animation can attract children's attention and it can engage children and maintain their motivation in specific contexts. Also, although live-action SAM videos are close to the real MCAST test, animations require less time and resources, such as a camera crew and specific location, to be produced. By identifying the role of different media types in digital stories on children's story experience, this thesis contributes to find the efficient and cost-effective media types of how to produce an engaging story using different multimedia technologies.

Therefore, these findings will demonstrate if the system could monitor child engagement levels successfully and identify 'disengaged' children. Automated engagement measurement reduces the need for so much time from trained assessors and ensures the quality of the data that will be used to make assessments, improving the efficiency of coding attachment evaluations. Meanwhile, when displaying digital story-stems to children, people without MCAST training, such as teachers, could administer the MCAST test to reduce the cost and involvement of fully trained assessors.

1.2 Research Questions

To summarise, this thesis focuses on developing a set of tools to measure children's engagement levels and investigating the effect of media types used in the digital story-stems on children's engagement levels. These tools are developed and validated through analyses of facial data from different age groups (children and adults). The main research questions for this thesis are:

RQ1: Can children's spontaneous facial expressions to be used to automatically measure engagement levels in digital story-stems?

RQ2: How do voice type and presentation type affect child engagement levels in digital story-stems?

1.3 Thesis Statement

This thesis focuses on measuring children's engagement levels in digital story-stems and investigating the effect of multimedia tools on creating a better and more engaging digital story. A set of measurements is developed and validated through analyses of children's spontaneous facial expressions when watching digital stories using different media and voice types. Results showed child engagement levels can be measured by certain facial measures and that the quality of storytelling voice to create stories and animated videos significantly affect children's level of engagement.

1.4 Thesis Structure

Chapter 2: Literature Review: This chapter describes various definitions of engagement in different contexts and defines the terminologies engagement and child engagement used in this thesis. The MCAST test, on which this thesis has based, is introduced and provides an initial scale of coding for children's engagement levels. The use of multimedia tools in designing digital story-stems is discussed and gaps are identified (RQ2). Finally, the benefits and disadvantages of engagement measurements are introduced to give a support of designing experiments to answer RQ1. Three kinds of measurement methods are used including the self-report measures, external observation and automated measures based on computer vision.

Chapter 3: An initial study of adult engagement measurements: This chapter describes a preliminary study for on measuring adult's engagement using their gaze behaviours. An analysis of the results shows the relationship between video quality and adult's engagement. This provides insight into how to design a method of measuring children's engagement and starts to answer RQ1.

Chapter 4: Child engagement measurements from facial data: This chapter builds on Chapter 3 and develops a set of tools that can measure child engagement levels while

watching digital story-stems. Data used in this chapter are focused on children's facial behaviours to answer RQ1.

Chapter 5: Designing an engaging digital story-stem: This chapter investigates the role of storytelling voice and animation vs. live-action recorded video as two multimedia types that can be used for engaging children in digital story-stems. It provides a better set of multimedia tools for designing the story-stems, which aims to answer RQ2.

Chapters 6: Discussion and Conclusions: This chapter discusses the answer to each research question, the limitations of the work in this thesis and give a final conclusion of this research.

Chapter 2 Literature Review

2.1 Outline

This review is structured into four parts. Section 2.2 describes various definitions of engagement in different contexts to show how it is explained and used. The second part (Section 2.3) introduces a psychological evaluation method – Manchester Child Attachment Story Task (MCAST) – on which the main work in the thesis is based. MCAST is a structured doll play methodology which uses short story-stems to assess the Attachment status of children. The extent to which children become engaged in the storytelling determines whether MCAST experts can assess children’s attachment status based on their behaviour during the test. Section 2.4 gives a description of multimedia tools for designing digital stories with the aim of making them more engaging. The last part (Section 2.5) presents three kinds of measurement methods that can be used to assess engagement including self-report measures, external observation and automated measures based on computer vision. From this review, the two thesis questions are then drawn out.

2.2 The Definition of Engagement

The term *engagement* is used in many different ways across many different research areas. Some of the key definitions will be provided from these different domains and discussed in this section to give background on the main concepts involved. These are used to inform the specific definition of engagement used for the rest of this thesis.

Engagement researchers suggest that there are two issues that need to be considered when designing experiments [54]. The first is the definition of engagement based on the different purposes of research. The second issue is to choose the proper methods that match the definition. This section provides various definitions of engagement in different fields, thereby discussing the commonalities and differences among the definitions.

2.2.1 Engagement in Human-Agent Interaction

Engagement is a key concept in investigating people's interaction with computer-based agents for the design and implementation of intelligent interfaces. Researchers have explored different definitions and meanings of engagement in this context.

When investigating the quality of the experiences with social robots, a common definition comes from Sidner *et al.* [82] who define engagement as “*the process by which two (or more) participants establish, maintain and end their perceived connection*”. The process of engaging with a robot includes initial contact, negotiating collaboration, checking that the other is still taking part in the interaction, evaluating staying involved and deciding when to end the connection. For example, an initial contact could be an interaction that a user may engage with a machine by moving into a specific range in which the machine could respond [70].

Using this definition, researchers have developed several kinds of robots to establish and maintain a face-to-face conversation with a person [3, 41, 59, 81, 92]. However, such conversations require more than just talking. It is entirely possible to build or build and maintain the engagement process without a single word being said. A person who engages with an agent without spoken language can depend on nonverbal behaviours, such as facial expression and gesture, to establish or maintain engagement [59]. Therefore, one important aspect to establish a natural conversational agent is that the system must monitor the user's nonverbal behaviours and estimate the engagement based on the behavioural information. Information collected using eye-tracking technology could help measure the “social connection”. For example, a robot should be able to receive and assess the eye gaze data from the human conversational partner as a listener's eye-contact could express his/her attention towards the conversation [39]. Ishii *et al.* have constructed a series of experiments to indicate that mutual gaze occurrence, gaze duration and eye movement distance provide optimal performance and can accurately estimate user engagement [39, 40, 42, 59]. It is also necessary to display the robot's facial expressions and gestures to signal that the robot is listening to the user. If a robot can detect whether the user is engaged in the conversation, then it can discover the objects of interest in the conversation as well as adapt its behaviour and communication according to the user's engagement levels.

Therefore, there is the general definition of engagement in Human-Robot Interaction (HRI) from Sidner *et al.* [82]. It is accompanied by a focus on process and is widely applied in the

design and evaluation of interaction with robots. Engaging with robots requires the perception of users' attention, which can be influenced by factors such as the effect of interaction distance on the visual cues. According to these applications between human and humanoid interface (e.g. robots), nonverbal behaviours can be obtained based on eye gaze, head gesture and facial expressions. These signals can be interpreted as direct features for measuring or predicting the engagement in face-to-face conversation. For example, not looking at the TV can be a good indicator of low engagement in the context of TV viewers [36]. The work in this thesis builds on these measurement methods to exploit features based on eye gaze and facial gestures to measure the level of engagement of a person who was watching a digital story. A more detailed description about measuring gaze and facial behaviours will be discussed in Sections 2.5.4 and 2.5.5.

2.2.2 Engagement in User-System Interaction

Besides engagement in HRI, some other interpretations of engagement are used in the more general context of user-system interaction. In this context, researchers often discuss the term "user engagement". An important contribution to this term comes from O'Brien and Toms [64] who describe engagement as: "*a quality of user experience with technology that is characterised by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, interest, and affect*". This definition positions engagement as users' experience – a component of human-information interaction. To assess the user perceptions, a post-experience questionnaire called the User Engagement Scale (UES) [63] was developed in four domains (online shopping, web searching, educational webcasting, and video games) to identify six factors: Perceived Usability (PUs), Aesthetics (AE), Novelty (NO), Felt Involvement (FI), Focused Attention (FA), and Endurability (EN) as shown in Table 2-2 (Section 2.5.1). This scale provides a self-report measurement method that gives a starting point for this thesis, in which engagement was measured by the depth of participation for each attribute depending on the interaction between the user and system during the experience.

Additionally, other researchers have described the term engagement with an appropriate prefix for the purpose of their particular studies in the context of HCI. Yu *et al.* [93] defined conversational engagement between users of a voice communication system that measures the commitment to interaction: "*user engagement describes how much a participant is interested in and attentive to a conversation.*" Bickmore *et al.* [11] recognised long-term

engagement between users and vocabulary-based systems as: “*the degree of involvement a user chooses to have with a system over time*”. These definitions explained how and why the specific applications attracted people to use them and gave several key attributes of engagement such as attention, interest and involvement.

O’Brien and Toms have provided a new definition of “user engagement” in the context of general user-system interaction and developed a scale called the User Engagement Scale (UES) to evaluate user experience in four different areas. This scale gives a starting point for designing a questionnaire of the self-report measure in this thesis. While metrics (e.g. facial expressions and eye gaze) aim to measure people’s behavioural engagement, a self-report measure could reflect users’ mental and cognitive states during the experience, both of which are crucial to engagement.

2.2.3 Engagement in Education

Student engagement has been a crucial topic in the field of education because it is a potential way to address low academic achievement, student boredom, student disengagement, and high drop-out rates. Several studies have investigated this [29, 33]. The National Research Council [60] indicated that increasing student engagement has been an explicit goal of many school and district improvement efforts, especially at the secondary level. Youths with high disengagement are less likely to graduate from high school and face limited employment prospects, so that to increase their risk of poverty, poorer health, and involvement in the criminal justice system. Therefore, teachers work to increase their student’s engagement because they know it is critical to student success. Increasing student engagement is not only related to traditional classrooms but also focused on other learning activities such as game-based learning, intelligent tutoring systems (ITS), and massively open online courses (MOOCs) [88].

Researchers have developed various definitions for describing student engagement. In these definitions and related concepts, the two most commonly applied definitions include the three dimensions from Fredricks *et al.* [28, 29] and the four dimensions from Appleton *et al.* [4]. Fredricks *et al.* [28, 29] proposed that student engagement can be characterised by behavioural, emotional and cognitive dimensions. The term *behavioural engagement* is typically used to describe the student’s willingness to participate in the learning process, e.g., attend class, stay on task, submit required work, and follow the teacher’s direction. It includes involvement in academic and social or extracurricular activities and is considered

crucial for achieving positive academic outcomes and preventing dropping out. *Emotional engagement* describes a student's emotional attitude towards learning, teachers and school. It is presumed to create ties to an institution and influence willingness to do work. *Cognitive engagement* refers to learning in a way that maximizes a person's cognitive abilities, including focused attention, memory, and creative thinking. It incorporates thoughtfulness and willingness to exert the effort necessary to comprehend complex ideas and master difficult skills.

There is some disagreement with these three dimensions of student engagement. For example, some teachers consider engagement as encompassing three interconnected dimensions: behavioural engagement, cognitive engagement, and relational engagement [22]. *Relational engagement* is said to be the most relevant to classroom management that promotes optimal engagement in school. Several researchers presented engagement in science learning with four dimensions including behavioural, emotional, cognitive, and agentic engagement [74, 83]. The concepts included in first three types of engagement overlap with constructs from previous studies. Besides these three, *agentic engagement* was firstly proposed by Reeve and Tseng and defined as students' constructive contribution into the flow of the instruction they receive [74]. It can modify (e.g. changing the level of difficulty) and enrich (e.g. make the task more enjoyable) students' learning activities, while behavioural, emotional, and cognitive engagement can only reflect the extent of students' reaction to learning activities. Appleton *et al.* [4] proposed an alternative framework based on previous studies [27, 29] that included four dimensions of student engagement: academic, behavioural, cognitive and psychological. The definitions of *behavioural* and *cognitive* engagement are similar to previous concepts from Fredricks *et al.* [28, 29]. *Academic engagement* consists of a subset of academic behaviours such as time on task, credits earned toward graduation, and homework completion. *Psychological engagement* refers to student's relationship with teachers and peers and feelings of identification or belonging.

Therefore, a common perspective across student engagement research has characterised it to include behavioural, emotional, and cognitive dimensions [28, 29]. These interpretations express engagement in more easily quantified ways as each dimension focuses on different components. Cognitive engagement is not used in this thesis because it focuses more on students' learning abilities such as memory and creative thinking, while this thesis focuses on the behaviours and emotions of story watching. According to this, the conceptualisations of engagement in this thesis will be based on the descriptions of behavioural and emotional

engagement. The term *behavioural engagement* will be used to describe participants' willingness to watch the digital stories, promising more objective measures using non-verbal cues such as head direction and gaze, smiles and gestures. Definitions of *emotional engagement* here tend to emphasise participants' experience including belonging, attitudes, and emotions. The sense of belonging will be measured by asking the extent to which the child participant feels like a part of the stories. Attitudes and emotions will be reflected through expressions of interest, boredom, and feelings.

2.2.4 Child Engagement in Preschool Classrooms

McWilliam and Bailey defined child engagement as: "*the amount of time children spend interacting with their environment in a developmentally and contextually appropriate manner*" [2]. Measuring children's engagement in pre-school classrooms is beneficial for understanding their behaviours in digital story environments in this thesis.

Educators indicated that increased engagement could reduce negative behaviours and promote social, physical, psychological and cognitive skills and abilities related to learning. To create an engaging classroom, there are three ways to promote children's engagement: changing the routine, children's expectations, and teachers' interaction styles [46, 52, 53, 72]. For example, physical environment affected children's engagement levels and specifically on increasing their engagement levels by designing a modified open classroom and using the developmentally appropriate materials [72]. Furthermore, researchers have demonstrated that children's engagement levels were affected by their age and disability status. Children with disabilities were likely to spend more time with interactively engaged peers than adults, and attentionally engaged with than children without disabilities; they spent more time passively non-engaged [53]. Teachers' interaction behaviours typically include prompt questions such as asking why and how questions, eliciting behaviours based on the children's interests and responses, a variety of non-intrusive strategies, such as modelling and time delay, and redirecting behaviours like stopping children and getting children to do something different from what they are doing [46]. Teachers' interaction style can provide information to expand on children's engagement and overcome children's general disengagement. Additionally, children were more engaged in classrooms when teachers addressed them individually than as part of a group [52].

According to these three aspects of constructing an engaging classroom, methods to measure children's engagement in different routines include rating the amount of time spent with

adults, peers and materials respectively. There are three rating systems to be discussed here: the Engagement Quality Observation System (E-Qual), the Children's Engagement Questionnaire (CEQ) for measuring child engagement behaviours, and the Child Caregiver Interaction Scale (CCIS) for measuring the teacher interaction style.

The E-Qual observational coding system was developed by McWilliam [52, 72]. Each child's engagement behaviour was coded by a level and a type. The E-Qual system includes four categories for measuring engagement in preschool classrooms: sophisticated, differentiated, focused, and unsophisticated. These focus on children's behaviours towards/with peers, adults, objects/materials, and self. Detailed definitions of the coding systems are shown in Appendix C [72]. This scale provides information on several engagement behavioural aspects that could be used in the context of story-stems. For example, the category called focused attention involves watching or listening to features in the environment for a duration of at least 3 seconds; it includes physical characteristics such as serious facial expression and subdued motor activity. In the context of this thesis, this would be watching the digital story-stems for a duration of at least 3 seconds, including attentional facial expressions.

The Children's Engagement Questionnaire (CEQ) [52] is an instrument for teachers to rate children's engagement based on their impressions of their abilities. It has four categories including not at all typical, somewhat typical, typical, and very typical, which gives a support of measuring engagement levels using external observation in four steps.

The last method is the Child Caregiver Interaction Scale (CCIS), to assess the quality of caregivers' interaction with children in care. The original CCIS was developed in 1989 [5] and revised by Carl [18] in 2010. In Carl's version, it is a 14-item instrument consisting of three domains. Each item uses a seven-point scale ranging from 1 (inadequate) to 7 (expanding) with clear description along the scales at 1, 3, 5 and 7. The first domain is the emotional, including 4 items – tone of voice/sensitivity, acceptance/respect for children, enjoys and appreciates children, and expectations for children. The second domain is the cognitive/physical, including 7 items – health and safety, routines/time spent, physical attention, discipline, language development, learning opportunities, and involvement with children's activities. The last one is the social domain, including 3 items – arrival, promotion of prosocial behaviour/Social Emotional Learning (SEL), and relationships with families. Several items are related to this thesis. For example, an engaging tone of caregivers' voice

emotionally expresses acceptance to children even in the predicament of a distressing story, which could be used as a standard for the storytellers' voice. In the cognitive/psychology domain, children's physical attention can be used to measure their engagement levels. A scale for physical attention is to get the eye contact between the child and the caregiver. Eye contact in the context of this thesis would take place where the child is watching the digital stories. Children's eye data could be captured and analysed using automated measures to understand their engagement levels.

A general definition of child engagement provided in this section. Although the main area of child engagement is the preschool classroom, this definition has also been used in other research. For example, researchers investigated child enjoyment and engagement while doing puzzles to find the design implications of tangible user interfaces [91]. The definition of child engagement in this thesis will be based on the definition from McWilliam and Bailey [2] for child engagement. The three methods (E-Qual, CEQ, and CCIS) provided a series of children's engagement behaviours and basic scales that could be used for external observation and story design in this thesis.

2.2.5 Narrative Engagement

This section focuses on a specific context – the story narrative. Story stem narratives are a reliable and valid method for observing children to find out how they think and feel about important relationships such as with their family [77]. The test used in this thesis, MCAST (see Section 2.3), is one instance of the story-stem approach.

Previous studies indicated that the extent to which people become transported, immersed and engaged in a narrative influence their subsequent story-related attitudes and beliefs. To understand the experience of engagement in narratives, Busselle and Bilandzic [16] proposed a mental model approach to explain the process in relation to narrative experiences. Mental models were constructed by a story reader or viewer to represent a story narrative through combining information (e.g. characters, environments, and situations) from the story with knowledge that originates in daily life and/or in specific topics related to the narrative. For example, Figure 2-1 shows an example of a child's mental model for a story related to a nightmare when she was listening to that story. In this figure, the child has a nightmare, which she is trying to escape by a fear of the devil in her nightmare. This image of the child's mental model displayed that she locates herself within the story, which means that she strongly emotionally identifies with the character. This is essentially empathy, an important

item for measuring narrative engagement [17]. As the story moves forward, these models are constantly updated.

The main application of the mental model approach in this thesis is to provide theoretical support for designing a story that broadly applies to media content. A hypothesis in this thesis is that the extent of people's narrative engagement differs due to different media methods of displaying stories to them, rather than the way that they process the information they receive. For example, compared to traditional oral storytelling, video provides visual and aural information for audiences to help them build a mental model for imagery.



Figure 2-1. An example of the mental model approach¹. The girl builds an image for a mental model that she has a nightmare when she was listening to a nightmare story.

Therefore, the mental model approach gives a support for using media content to design the digital MCAST story-stems on a practical level. Digital MCAST stories could help children build their mental model for imagery to improve their engagement and comprehension to the story. Displaying digital MCAST stories as a key step of automating the use of story-stems approach could reduce the cost and time required for test administrators. More details about how to design the digital story-stems will be described in Section 2.4. Furthermore, Busselle and Bilandzic [17] also developed a scale based on the mental models approach for measuring narrative engagement with four dimensions: narrative understanding, attentional focus, emotional engagement, and narrative presence. Items related to the four dimensions could be used for designing a self-report questionnaire for measuring children's engagement in this thesis, to be discussed in Section 2.5.1.

¹ <https://www.dreamstime.com/stock-photos-girl-having-nightmare-whilst-sleeping-hand-drawn-picture-child-dreaming-illustrated-loose-style-vector-eps-available-image32518253>

2.2.6 Discussion

The section introduced the term *engagement* used in five different research areas. Different motivations and purposes are the key aspects of defining the term *engagement*. To date, analysis of the definition of engagement has been limited, and difficulties understanding and adopting the term persist, a topic of concern for many within the community.

Previous studies have used many different definitions of engagement, as shown in Table 2-1, where engagement is positioned at the level of individual conscious experience that is described as a process of perceived connection (Sidner), a quality of user experience (O'Brien), or a complex concept comprising several dimensions (Fredricks, Appleton, McWilliam, and Busselle). These definitions are crucial to the value of engagement as a concept. According to these definitions, engagement is a complex concept reflected in several modalities including face and body language, speech and physiology, which can be measured by subjective and objective measurement methods. Behavioural engagement generally promises more objective measurement based on users' behaviours while emotional engagement focuses on user experience, attitudes and emotions.

Reference	Participants	Research area	Definition	For this thesis
Sidner <i>et al.</i> [82]	Between participants	Human-agent interaction	<i>Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection.</i>	Nonverbal behaviours (e.g. eye gaze, head gesture and facial expressions) can be interpreted as direct features for measuring or predicting the engagement.
O'Brien and Toms [64]	A user towards the system	User-system interaction (e.g. online shopping, web searching, educational webcasting, and video games)	<i>Engagement is a quality of user experience with technology that is characterised by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, interest, and affect.</i>	A scale called User Engagement Scale (UES) has been developed to evaluate user experience in different fields, which gives a start point for designing a questionnaire of the self-report measure.

Fredricks <i>et al.</i> [28, 29]	Students	Education	<i>Student engagement has been characterised with behavioural, emotional and cognitive dimensions.</i>	The term <i>behavioural and emotional engagement</i> will be used to describe the conceptualisations of engagement.
Appleton <i>et al.</i> [4]	Students	Education	<i>There is an alternative framework based on [27, 29] that included four dimensions of student engagement: academic, behavioural, cognitive and psychology.</i>	<i>Behavioural engagement</i> promises objective measures using non-verbal cues. <i>Emotional engagement</i> focuses on subjective measures that emphasise participants' attitudes and emotions through expressions of interest, boredom, and feelings.
McWilliam and Bailey [2]	Children	Education and family interaction	<i>Child engagement is the amount of time children spend interacting with their environment in a developmentally and contextually appropriate manner.</i>	Providing a definition and several coding systems for measuring child engagement.
Busselle and Bilandzic [17]	Participants	Narrative	<i>Narrative engagement has been distinguished among four dimensions of experiential engagement in narratives: narrative understanding, attentional focus, emotional engagement, and narrative presence.</i>	A scale called Narrative Engagement Scale from has been developed to foster understanding of the experience of engaging with a narrative, which gives a support for designing a questionnaire of the self-report measure.

Table 2-1. Definitions of engagement across different research areas.

2.3 The project focus – Manchester Child Attachment Story Task (MCAST)

Story stem narratives are a reliable and valid method for observing children to find out how they think and feel about important relationships such as with their family [77], and have made significant contributions to understanding attachment with caregivers [14, 15, 32]. Attachment with caregivers (typically the mother) is one of most important aspects of young children's relationship functioning. There are several methods to assess middle-aged (4-8 years old) children's representation of attachment by using story-stems and doll play completions such as the MacArthur Story Stem Battery (MSSB) [15], the Attachment Story Completion Task (ASCT) [14] and the Manchester Child Attachment Story Task (MCAST) [32].

MCAST is a structured doll play methodology, using short story-stem vignettes to assess child attachment representations in relation to a specific primary caregiver [32]. It has good inter-rater reliability, stability of attachment patterns and has been validated against other attachment measures including the Adult Attachment Interview (AAI) and Separation Anxiety Test [30].

In the original MCAST, an assessor shows five story-stem vignettes to the child, using a dolls-house. An initial breakfast vignette represents an introduction to the procedure and a non-attachment comparison. There are then four attachment-related vignettes in a situation of specific mild 'distress' with the caregiver. The stories include scenarios around: Nightmare, Hurt Knee, Illness, and Shopping. In the nightmare story, a child doll (whose mother doll is in another bedroom) awakes at night alone with a nightmare; in the hurt knee story a child doll (whose mother doll is in the dolls-house) is represented as falling and hurting her knee while out in the garden; the illness story shows a child doll developing a sore tummy while watching a favourite TV program; and in the shopping story, the child doll suddenly finds him/herself lost and alone while shopping with the mummy doll in a large crowd.

During the MCAST test, the child is asked to listen each story-stem vignette and then act out what happens in the rest of the story with symbolic dolls. The MCAST setup was shown in Figure 1-1 (Section 1.1). The way the child completes the story and their behaviour during the test provides the cues necessary to assess their Attachment status. The test takes between

20 and 30 minutes to administer and from one to two hours to code from videotape, depending on the complexity of the material and the child's attachment status.

For each of four 'distress' scenarios, there is an induction phase where a child is given the beginning of a story by an assessor using two dolls. In this phase, the story examiner 'amplifies' the intensity of anxiety and distress represented in the child doll (e.g. the child wakes at night alone with a nightmare), prompting the child to resolve the scenario during the story completion phase. The child is observed to see if he/she is engaged at the emotional level by the predicament shown in each distressed story until he/she is able to play with the dolls to complete the story spontaneously. The aim of this phase is to bring children into a deep engagement with the mildly stressful story to bring out their mental representation of attachment to their caregiver [32].

Engagement in the induction phase is important as it means that children focus on attending to the play and materials, are not distracted by other things, and feel empathy with the dolls and characters in the story. If the child is not engaged, the test cannot be administered correctly, and results will not be analysable to give Attachment status. Engagement is measured by a trained assessor's observation of facial expressions, using the standard MCAST protocol. In the protocol, the engagement scale is a general schema ranging from 1 to 9 with clear descriptions along the scale at 1, 3, 5, 7, and 9, and these 9 levels are grouped into four ranges: normal/optimal range (score 7-9), borderline (score 5-6), abnormal scores (score 3-4) and seriously abnormal scores (score 3 and below). From the MCAST protocol:

1. Impossible to engage. Either overactive, distractible and unable to focus or extremely passive.
- 2.
3. Examiner/Storyteller has to work much harder than usual but still cannot keep develop the child's engagement successfully.
- 4.
5. Good enough to proceed to the next phase but still somewhat problematic and examiner has to work quite hard to initiate/maintain engagement. Below 5 the observer will not be able to proceed with the interview. Above 5 the interview can proceed.
- 6.
7. Good quality engagement by the end. Examiner only has to work slightly to maintain engagement.
- 8.
9. High quality full engagement from the beginning. Immediate engagement with play materials and intense active interest in the story. Deepening concentration as vignette proceeds.

From this scale, if an assessor thinks that a child's engagement score is lower than 5, which means the child is not engaged by the predicament shown in each story-stem, the assessor will stop the test and the child cannot be assessed for Attachment status based on their story and behaviour during the activity.

Unfortunately, conducting MCAST assessments is expensive and time-consuming. Examiners must attend high-cost courses followed by lengthy reliability training to be certified to perform MCAST [55]. Furthermore, the efficiency of MCAST assessment is limited by the number of children that they can reach. Trained assessors must spend time observing children's facial expressions from video recordings of the administration of the test and rating the child's engagement levels, which takes a long time [49]. This means that few children are tested. Early diagnosis of attachment problems makes treating the condition more straightforward. If untreated, it can lead to many problems later in life, from aggressive behaviour to cardiovascular disease [38].

There are two studies [55, 87] to reduce the time and cost required for MCAST administration and assessment. Minnis *et al.* [55] have developed the CMCAST, a computerised version of MCAST, which can be used on any personal computer. Story-stem vignettes are represented by animations on the computer as shown in Figure 2-2. Children are asked to watch the vignettes shown on the screen then take over the mouse and complete each story by speaking to the computer and moving the dolls presented on screen. A webcam records their audio and video. The use of CMCAST reduces the intensive involvement of trained MCAST administrators as it does not require full MCAST training to use it, so that costs are lower. Moreover, children's story and behaviours during the test are recorded and are stored automatically to reduce the chance of data loss. MCAST assessors can download the data for rating children's Attachment patterns.

CMCAST demonstrated that displaying the story-stems on screen was possible and could be used for successful MCAST measurement. However, an administrator still had to be present to assess children's engagement during the CMCAST test, using the MCAST protocol. Some issues around engaging the children were identified [55]: each child's engagement was labelled as "yes/ no" by an administrator while the child was watching the CMCAST stories and 16% (14/86) of children were labelled as insufficiently engaged. However, the authors have not explained why children were disengaged during the CMCAST administration. For instance, they did not investigate different media types and how they could affect children's

engagement; they only used simple animations controlled by a mouse. This thesis will look at the effects of different media types to see which is the most engaging. In addition, the CMCAST still needs the same training as MCAST for assessment. Attachment status was still coded by a trained assessor's observation of children's story and behaviours, using the MCAST protocol.



Figure 2-2. CMCAST computer interface [55]. Story-stems are represented by animations on the computer. Children are asked to watch the vignettes then complete each story by speaking to the computer and moving the dolls presented on screen using the mouse.

CMCAST was taken further a system called the School Attachment Monitor (SAM) [78, 87], which is designed to automate attachment assessments fully by administrating the MCAST test and automatically classifying the resulting attachment data (Figure 1-2 in Section 1.1). In this case, no human input would be required.

The aim of SAM is to develop a computer-based tool which can measure parent-child attachment across the population in a cost-effective way. It aims to make large-scale attachment screening possible by reducing time and costs required for MCAST assessment. The approach of SAM consists of automating the key steps of MCAST to 1) reduce the time needed to administer the test (higher efficiency); 2) reduce the time taken to assess the results (lower costs). Using SAM, MCAST can be administrated by non-experts, such as teachers in a classroom, as it can automatically administer the tests.

SAM is made of two pieces of software for administrating the assessment and collecting data. There are four 'distress' story-stem vignettes taken from MCAST by changing some details and the scripts of the vignettes show as following:

The 'Nightmare' story-stem:

In this story, it's in the middle of night and the mummy doll and the child doll are in their beds fast asleep.

Everything is very dark and very quiet.

Then suddenly the child doll wakes up.

And he says "Ooooh... I've had a horrible dream ooh... a horrible horrible dream...". And he starts to cry and he says "I was so scared... oowwwwww it was a terrible terrible dream..."

Now you show me what happens next...

The 'The Hurt Knee' story-stem:

In this story, it's daytime and the mummy doll is inside the house - let's say she is cooking in the kitchen.

Child doll is outside playing in the garden.

Look! He is playing hopscotch. So the child doll jumps, and jumps, and jumps...

And he jumps higher and higher and it's almost the end! ... And Oh no! The child doll slips in a puddle!

"Ooooh..." he cries "I hurt my knee... and it's bleeding ... oowwwwww my poor knee..."

Now you tell me what happens next...

The 'Illness' story-stem:

In this story, it's daytime and the child doll is at home watching TV – What's your favourite TV programme?

Child doll is watching that programme.

And the mummy doll is in the next door. Let's say she is in the kitchen.

Suddenly the child doll has a pain in his tummy.

And it gets worse!

The child doll cries "Ooooh... I've got a pain in my tummy... oowwwwww it's getting worse... oowwwwww... a horrible pain..."

Now what happens next?

The 'Shopping' story-stem:

In this story, it's daytime and the child doll and the mummy doll are out and about – they are going shopping.

Here they go into the shopping centre and there are crowds of people around so they have to hold on tight to each other.

They look in this shop here. And then they go to the shop there.

The child doll wants to look in this shop...

The child looks around and he finds he can't see his mummy.

The child doll tries to find in this shop and he tries to find there... There are all the people around but mummy is nowhere to be seen.

The child doll feels very scared and he cries "Ooooh...where's my mummy? Where's my mummy? "

Now you show me what happens next...

During the SAM administration phase, there is no story administrator to show the vignettes to children; children are guided through the story-stem vignettes on screen. The detailed movements of children are captured in real time on a laptop from a web camera and sensors in the dolls. The SAM system itself does not detect the child's engagement. Ensuring that children are engaged in SAM is vital as the system will collect poor data that cannot be used for assessing their Attachment status if they are not engaged. The aim of the work in this thesis was to understand child engagement so that engaging presentations of the story-stems could be created, along with ways of detecting whether the children were engaged. The latter aspect is done using the data collected from the SAM system.

Automated engagement measures will reduce the time and effort of constructing the SAM coding system due to the reduction in poor-quality data collected. If a child is identified as 'disengaged', recordings of this child cannot be used to assess the attachment status based on their story and behaviour during the activity. Next section will be described how to design more engaging story-stem vignettes using multimedia tools.

2.4 Multimedia Tools for Storytelling

Multimedia technology as a digital story design tool includes graphics, animation, text, recorded audio narration, video and music. The combination of graphics, animation, text,

recorded audio narration, video and music to create stories are called digital stories. Story-stem vignettes are represented on a screen by the movements of two dolls narrated by a human storyteller's voice. As this thesis is focused on designing an engaging digital story-stem, this section describes multimedia tools including animation, live-action video and storytelling voice, which are key aspects of digital stories. A good voice in a digital story makes audiences understand the story line and really "get into" the story. Audiences may not be consistently engaged where a narrator's voice does not fit the story line or a narrator's voice has a flat tone [86]. A detailed description of these three media types is given in the following subsections and a study to illustrate the effect on child engagement of different media in the digital story-stems will be discussed in Chapter 5.

2.4.1 Animation vs. Live-action video

Animation and live-action video are two important types of presentation. Live action² involves "real people or animals, not models, or images that are drawn, or produced by computer". Animation³ is "a film for the cinema, television or computer screen, which is achieved by a motion picture that is made from a series of drawings, computer graphics, or photographs of inanimate objects (such as clay, puppets) and that simulates movement by slight progressive changes in each frame". Both of these are possible to use in story-stem vignettes and this thesis will investigate which is most effective.

Live-action videos have been used in movies, games and marketing for attracting audiences. For example, live action is a great tool for connecting with customers in business, such as demonstrating a tangible product and displaying a consulting firm or a restaurant. It gives audiences an effective emotional connection when telling the story. For designing digital stories, SAM, as discussed in Section 2.3, uses live-action videos for displaying the MCAST story-stems. Live-action videos in SAM involves the video recording of a storyteller holding two physical dolls and performing the vignettes. The live-action video in this thesis was based on the SAM videos and children's engagement was measured while watching these. However, this may not be the best way of engaging the children, as other methods of presentation are available.

² <https://dictionary.cambridge.org/dictionary/english/live-action>

³ <https://www.merriam-webster.com/dictionary/animated%20cartoon>

Several studies have investigated the role of animation on engaging people in various fields, including entertainment, commercial, educational and personal purposes. For example, the use of animations has become widespread in education since the early 1980s as many animation designers recommended that animations can help communicate complex ideas more easily [1]. “Animation is the language of childhood,” said John Martin, Reallusion⁴. It has been widely used in the area of education. With properly designed and implemented animations, children are able to enhance their learning. Firstly, animations can be used to make exciting and fun narratives in which education and training can easily be incorporated. Secondly, animation can engage children and sustain their motivation as an effective learning tool. This affective animation training portrays interactive, creative, fun and motivational activities instead of comprehension of academic subjects. Children in these activities are fascinated by animation and animated stories and enjoy the opportunity to create their own. Animation can attract children’s attention on the screen during the storytelling processing [37]. These studies suggest that animation might be a good alternative for the design of the MCAST story-stem vignettes in this thesis. Once children’s attention is captured, the distinctive objects (i.e. the symbolic dolls shown on screen) may bring children into a deep engagement with the MCAST story-stems so that children will be not distracted by other things and feel empathy with the dolls and characters in the story.

Both live action and animation present their own pros and cons for digital story design. Live action involves the filming of storytellers and the physical dolls, making the digital story close to the administration of the real MCAST test. The only drawback is budget. Live-action videos require more time and resources than animation, including camera crew availability and the need for specific locations for recording. For animation, the movements of two symbolic dolls and a storytelling voice are needed, which reduces the costs required. There are many off-the-shelf software tools to support animation construction that can be used. Moreover, although animation production is time-consuming, it would be a good way to display a story if the script is solid because details of animation can be easily revised⁵.

Therefore, Chapter 5 reports an experiment to compare animation against live action for the presentation of the MCAST story-stems. The focus of this study was to determine if children’s engagement levels perceived in an animated sequence version of MCAST stories were higher than a live-action video version of the same content. These findings have

⁴ <https://www.reallusion.com/education>

⁵ <http://www.toddalcott.com/screenwriting-101-animation-vs-live-action.html>

significant implications for the use of automating the story-stem approach as engaging children in the stories is vital.

2.4.2 Storytelling Voice

The emotional speaking style of the storytelling voice is an area of great interest across a variety of fields. Human storytellers use their voice in various ways, such as making special sounds and using prosody to convey emotion, to capture the audience's attention and create an engaging listening experience for audiences [84]. In digital storytelling applications, stories are told on a computer. To capture the audience's attention and keep them engaged with the story, two factors of a storyteller's voice are important: voice gender and voice expressiveness.

Voice gender

Human voice typically can be described using a number of unique elements: gender, pitch, age, timbre and intensity. A professional storyteller can work on his/her voice to increase/decrease its pitch and timbre to invoke emotions in audiences. However, gender is a given. The pitch of the voice, also called fundamental frequency, is defined as the "rate of vibration of the vocal folds"⁶. Females tend to have higher voices because they have shorter vocal cords. The pitch of voice is an integral part of the human voice to be used to distinguish male and female voices. For example, a typical adult male will have a fundamental frequency of voice pitch from 85 to 180Hz while a typical adult female from 165 to 255Hz [85].

The difference of the roles of male and female storytellers was demonstrated mostly in folklore studies since the beginning of the 20th century to decide the most appropriate gender of a storyteller's voice [44]. For example, audiences could be more easily engaged with a female storyteller's voice when the female storyteller tells a heroine tale because she can demonstrate the story in a woman's point of view, such as more detailed depiction on the female life or women's daily activities. Another example of the use of voice gender is Siri⁷, Apple's voice-activated virtual assistant. Siri's first narration voice was female in the US because people generally find women's voices more pleasing than men's voices and Nass

⁶ <https://www.yorku.ca/earmstro/journey/resonation.html>

⁷ <https://edition.cnn.com/2011/10/21/tech/innovation/female-computer-voices/index.html>

suggested that this preference might even start from birth because babies attend to a female voice more than a male one.

According to this preference of human voice gender, one purpose of this thesis is to investigate the role of voice gender on children's engagement levels. In the original MCAST, administrators may be either male or female and all administrators are certified to perform the MCAST test well. However, since the MCAST stories are displayed on a screen in the SAM test, identifying the role of female and male storytellers in MCAST stories will help researchers create a more engaging digital story for children. Since the MCAST story-stems are attachment-related vignettes in a situation of specific 'distress' with the caregiver (typically the mother), a hypothesis here is that children would be more engaged in a female storyteller's voice than a male's voice, as a female storyteller can demonstrate the story like a mother. An attractive digital MCAST story would improve children's engagement levels so that more valid and reliable data could be collected for assessing their attachment status.

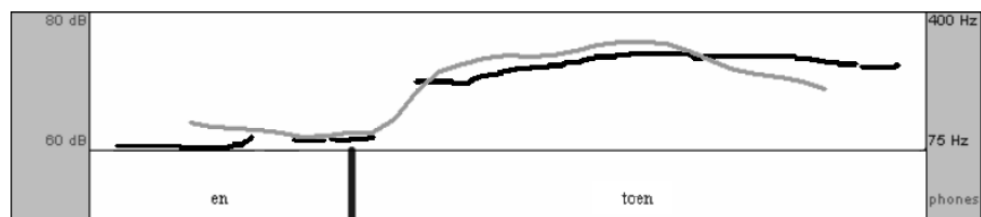
Voice expressiveness

Voice expressiveness has been studied in the area of the emotional expressiveness of a robot's speech. Many storytelling applications for young children have used social robots with a computer-generated text-to-speech voice in a language learning environment. These storytelling applications aim to deliver an equally engaging listening experience as that provided by a human storyteller. For example, Kory Westlund *et al.* [45] focused on the effect of the expressiveness of a robot's voice on children's engagement and learning. They observed that children's facial signs of higher emotional engagement and concentration when listening to an expressive voice rather than a flat robot voice, where this expressive voice was generated from the flat voice by modifying the prosodic parameters.

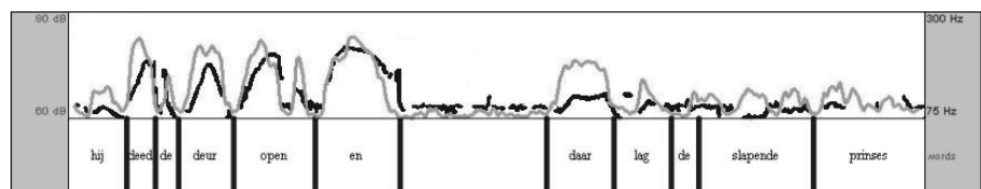
Although the storytelling speaking style can be achieved by modifying the prosodic parameters produced by a text-to-speech system, there is no synthetic speech system that can reach the full expressiveness of a human storyteller's speech. For MCAST, there are no computer-generated voices that can currently imitate the dynamic, expressive range of the voices of MCAST administrators because every storyteller must attend lengthy reliability training. A key question for this thesis is how different expressiveness in human's voice affects engagement. A study in Chapter 5 concentrates on the expressive style and flat style of human storytelling voices to investigate which is most effective for children to be used to

create a digital story-stem and understand the effects of voice attributes, such as pitch, on expressiveness.

The expressive storytelling style to be studied is to express suspense in the story. Every MCAST story-stem has a ‘distress’ situation and creating suspense makes the predicament in the story more stressful to engage children. There are two types of suspense: sudden climax and increasing climax [68, 84]. *Sudden climax* is an unexpected dramatic moment in the story, such as a startling revelation or a sudden, momentous event. It is typically announced by a dramatic increase of intensity and pitch on a keyword like ‘then’, ‘suddenly’, and ‘but’. The second type of suspense is *increasing climax* where the dramatic event is expected in advance. While approaching the climax, the storyteller heightens the suspense by a gradual increase in pitch and intensity, accompanied by a decrease in tempo. Figure 2-3 shows the two types of climax. The upper one illustrates a sudden climax in a fragment from the story of Bluebeard: ‘Her eyes had to get used to the darkness, and then ...!’ This climactic event was announced by a steep increase of intensity and pitch on the keyword (‘then’) introducing the climax. The bottom graph shows a fragment with an increasing climax from the story of Sleeping Beauty: ‘He opened the door and... there was the sleeping princess.’ The time domain for the increasing climax is split up into two parts, both typically spanning a clause. The first part builds up the expectation and ends with the key word announcing the revelation (e.g., ‘He opened the door and then –’) and the actual revelation takes place (‘– there was the sleeping princess’) in the second part [84]. This event was announced by a decrease of intensity and pitch, and a clear pause between en (‘and’) and daar (‘there’).



Sudden climax intensity and pitch. (Translation: ‘and then’)



Increasing climax intensity and pitch. (Translation: ‘He opened the door and... there was the sleeping princess.’)

Figure 2-3. Two types of climax in a story [84]. The upper one shows a sudden climax while the bottom one contains an increasing climax.

Theune *et al.* [84] developed rules for converting synthetic speech to expressive style speech. They compared prosodic features like pitch, intensity, tempo and pause duration of sample sentences recorded in two ways (neutral vs expressive) and some of them include a sudden climax or an increasing climax. They found that storytelling speech has more dynamics in prosodic parameters compared to neutral speech. The rules provided the optimal range of prosodic parameters (i.e., the value of pitch and intensity increase with respect to neutral speech), these were used to compare between the flat and expressive styles of human voices in this thesis to find the differences in child engagement. The differences among the voice parameters between the neutral and the expressive style of synthetic speech were formulated in the following rules:

(1) Pitch (fundamental frequency): In prosody, it is seen as the direct expression of intonation. The pitch contours of storytelling speech have rising or falling patterns (or both) with respect to neutral speech [68, 84]. Also, the mean value of pitch of expressive storytelling speech is higher or lower that compared to neutral speech [68, 84]. Within the time domain $[t_1, t_2]$, the pitch is increased gradually based on the observation in rise-fall patterns of the pitch contour. The pitch value of neutral speech is found to increase between 30-60Hz relative to the storyteller's average pitch on the accent syllables of sample sentences. The best value is 40Hz. The rules used to express a sudden climax is the pitch value is constant and a significant increase at the keyword for announcing the climax. The best value for pitch increase at the keyword is 80Hz. While expressing an increasing climax, the sample sentence should be divided into two parts including an expectation part from the beginning to the keyword $[t_1, t_2]$ and a revelation part $[t_2, t_3]$. There is a gradual increase of pitch in $[t_1, t_2]$; there is an initial pitch increase of 25Hz at t_1 and the pitch value gradually increases to 60Hz in $[t_1, t_2]$ for the accented syllables. In $[t_2, t_3]$, pitch value gradually decreases to its normal value.

(2) Intensity: Intensity is the correlate of physical energy and the degree of loudness of a speech sound [68]. In general, accented syllables in storyteller speech have a relatively higher intensity than neutral speech. Within time domain $[t_1, t_2]$ of accent syllables, previous research indicated that the intensity of storyteller speech increases between 2dB and 6dB relative to the average intensity of the neutral speech and found 2dB as the best value [84]. In a sudden climax, intensity is strongly increased at the keyword but then gradually decreases to its normal value. The initial intensity increases of 6 and 10dB relative to the speaker's average intensity. The best value for performing a sudden climax was 6dB.

In an increasing climax, the sample sentence should be divided into two parts including an expectation part from the beginning to the keyword $[t_1, t_2]$ and a revelation part $[t_2, t_3]$. An intensity increase of 10dB is constant across $[t_1, t_2]$ and a gradual decrease to its normal value across $[t_2, t_3]$. In addition, there should be a pause at t_2 , just before the revelation of the climactic event.

Besides pitch and intensity, there were also other vocal attributes but will not be used in this thesis. For example, tempo, also called the speaking rate, usually uses in emotionless stories, so that it was not suitable for comparing the voice expressiveness. The above rules will be used to check if there were differences across these storytelling voices. Therefore, this thesis will investigate the most suitable human storytelling voice for creating an engaging MCAST story, but the use of this work is not limited to MCAST story generation. It can also provide an expressive storytelling style that can be used to generate storytelling voices and resynthesise existing stories.

2.5 Measurement Methods

This section describes how to choose methods could be used for measuring engagement that match the definitions provided in Section 2.2.6. There are two broad types: subjective-oriented and objective-oriented measures. Subjective measures focus on recording a user's perception by using self-report measures (e.g. questionnaire, interview) to support users to express their attitudes, feelings, beliefs or knowledge about a subject or situation. Objective-oriented measures generally aim at searching for actionable data and each objective measure tends to target a very specific aspect of engagement instead of addressing a range of variables. There are five considerations of objective measures for engagement: the subjective perception of time, follow-on task performance, physiological sensors, online behaviour, and information retrieval metrics [6]. This literature review features three types of both approaches to measurement: questionnaires of self-report measures, external observation and physiological measures.

2.5.1 Self-report Measures

The self-report represents a robust, efficient and easy to implement approach for collecting valid, reliable data for assessing engagement in multiple areas such as a video game-based environment and education [51, 56, 61, 89]. Researchers have suggested that the self-report measure differs from objective-oriented measures as it provides a participant's perspective

of a system based on his/her cognition, emotion and memory to help researchers understand the participant's engagement [56, 61]. Approaches to study engagement with subjective measures include questionnaires, interviews, think-aloud protocols and other forms of self-reports. Since a questionnaire instrument is the most commonly-used technique for measuring engagement in prior research [23], there are two questionnaires to be discussed including the User Engagement Scale (UES), the Narrative Engagement Scale (NES) and an instrument called Intrinsic Motivational Inventory (IMI).

O'Brien and Toms have defined engagement in the context of user-system interaction and constructed the User Engagement Scale (UES) [61] as a post-experience questionnaire for assessing user engagement in four domains (online shopping, web searching, educational webcasting, and video games). This survey provided a conceptual model of user engagement in the context of HCI and validated six subscales: *Perceived Usability (PUs)*, *Aesthetics (AE)*, *Novelty (NO)*, *Felt Involvement (FI)*, *Focused Attention (FA)*, and *Endurability (EN)* as shown in Table 2-2. There are 31 items used for investigating online shopping experiences as displayed in Appendix A. Wiebe *et al.* [89] then extended research of O'Brien and Toms for developing a self-report instrument of engagement in computer and game-based environment. They revealed four subscales – *Focused Attention (FA)*, *Perceived Usability (PU)*, *Aesthetics (AE)*, and *Satisfaction (SA)* – as compared with the six in previous research. The fourth factor, satisfaction (SA), is a combination of items from the original Endurability (EN), Novelty (NO), and Felt Involvement (FI) subscales. They found that a self-report instrument with four subscales provides a better result than the model defined as six subscales in video game-based environment.

From the previous research, a crucial point of designing a questionnaire for measuring engagement is context-dependent. Since this thesis focuses on the level of child engagement in digital story-stems, the questionnaire aims to interpret engagement as “a generic indicator of a story viewer's state around the particular distress represented in the story-stem”. Aesthetics and Focused Attention taken from O'Brien *et al.*'s research [62] were pertinent to this thesis. Items related to the two scales were modified to fit the digital story-stem environment to indicate children's attitudes towards the content and different media types of the story-stem based on a 1 to 5 Likert scale. For example, one item of Aesthetics from the original UES was ‘I liked the graphics and images used on this shopping website’. The modified one for this thesis is ‘I liked the voice used on this story.’.

Subscale	Description
Perceived Usability (PU)	Both affective (frustration) and cognitive (effort) aspects of use of the system; Users' perception of estimated time spent on task.
Aesthetics (AE)	Visual beauty or the study of natural and pleasing (or aesthetic) computer-based environments
Novelty (NO)	Variety of sudden and unexpected changes (visual or auditory) that cause excitement and joy or alarm; Features of the interface that "users find unexpected, surprising, new, and unfamiliar"
Felt Involvement (FI)	Users' feelings of being drawn in, interested, and having fun during the interaction.
Focused Attention (FA)	The concentration of mental activity; concentrating on one stimulus only and ignoring all others; Focused concentration, absorption, temporal dissociation
Endurability (EN)	Holistic response to experience, likelihood of remembering an experience.

Table 2-2. Six attributes of the User Engagement Scale (UES) [63].

Meanwhile, clarifying the experience of engaging with a narrative provides a motivation of measuring narrative engagement in a theoretically meaningful way. Based on the mental models approach, Busselle and Bilandzic have developed a scale for measuring narrative engagement comprising four dimensions: *narrative understanding*, *attentional focus*, *emotional engagement*, and *narrative presence* [17]. Table 2-3 shows the descriptions of each dimension and related items used for developing the Narrative Engagement Scale (NES). Full items are shown in Appendix B. One point in the subscale of emotional engagement is that it concerns emotions that viewers have with respect to characters in the story, either feeling the characters' emotions (empathy), or feeling for them (sympathy). Sympathy differs from empathy because the audience member does not feel the same emotion as the character. In this thesis, children's emotional engagement would be measured using empathy, where they can feel with the child doll's emotion.

Since an important aspect of engagement is emotional, *distraction* and *empathy* as two main items will be used in this thesis to investigate the extent of children's attention and emotional engagement in the story-stems. Besides emotional engagement, *narrative comprehension* is also a necessary subscale that measures children's mental models within the story. It gives a theoretical support to the items of empathy and aesthetics (taken from the UES) discussed in Section 2.2.5. So far, there are five subscales to be used for designing the questionnaires including *Aesthetics* and *Focused Attention* taken from the User Engagement Scale (UES),

Story Comprehension, *Attentional Focus* (the item – distraction) and *Emotional Engagement* (the item – empathy) taken from the Narrative Engagement Scale (NES).

Dimensions of engagement	Description	Related items
Narrative comprehension	Narrative comprehension requires a viewer or reader locate him or herself within the mental model of the story.	Narrative realism; Cognitive perspective taking; Ease of cognitive access;
Attentional focus	Attentional focus means that a fully-engaged viewer should only be aware of attention shifts or re-attention rather than aware of focused attention.	Distraction;
Emotional engagement	Emotional engagement focuses on feeling with the characters' emotions (empathy) or feeling for them (sympathy), but not necessarily to any specific emotion.	Empathy; Sympathy;
Narrative presence	Narrative presence is the sensation that one has left the actual world and entered the story.	Narrative presence; Loss of self-awareness; Loss of time;

Table 2-3. Four dimensions of measuring narrative engagement [17].

Although O'Brien *et al.* [62] and Busselle *et al.* [17] investigated different areas to develop a standardised engagement questionnaire, different demographic groups may have different engagement characteristics. While usability may be a crucial property for adults, interest and enjoyment may be important characteristics for engaging children. Ryan proposed the development of an intrinsic motivational inventory (IMI)⁸ instrument to measure children's subjective experiences related to enjoyment and interest in experimental tasks. The IMI tool assesses the levels of interest/enjoyment, perceived competence, effort, value/usefulness, feeling pressure and tension, and perceived choice of children while performing an activity. It has been widely used in many studies because of the ease of customisation. Karimi *et al.* [43] revised the IMI instrument to measure children's playing learning experience in four subscales including interest and enjoyment, perceived competence, feeling pressure and tension, and perceived choice. Xie *et al.* [91] used the revised IMI instrument to investigate the relationship between the interface style and children's enjoyment while doing puzzles. According to the previous studies, children's interest to this task was as an important

⁸ <http://selfdeterminationtheory.org/intrinsic-motivation-inventory/>

subscale, which will be used to design the questionnaire because children's attitudes towards the task can reflect their engagement.

Subscale	Description
Aesthetics	Visual beauty or the study of natural and pleasing (or aesthetic) computer-based environments
Distraction/ Attentional focus	The concentration of mental activity; concentrating on one stimulus only and ignoring all others; Focused concentration, absorption, temporal dissociation
Empathy	Feeling with the character's (the child doll) emotions?
Story understanding	Story understanding is a sign of a child locate him or herself within the mental model of the story.
General attitude towards this task/Interest	Feelings of being interested and having fun during the story watching.

Table 2-4. Five subscales used for designing the questionnaire for children in this thesis.

Therefore, this thesis aims to design a questionnaire for children to investigate their mood state particularly around the distress situation in the context of watching the story-stems to interpret their engagement levels. That questionnaire was designed that was mainly based on the UES, NES and IMI. Since both the UES and NES have the subscale of focused attention, it would be combined. There are five subscales used for designing a questionnaire to measure child engagement: *empathy*, *distraction/attentional focus*, *story understanding*, *aesthetics*, and *interest/general attitudes towards the task*, as shown in Table 2-4. Moreover, as questionnaires usually invite closed-ended responses, there are two open-ended questions at the end, which allow children to describe their experiences and attitudes. For example, a question like "one story character's feeling at the end" in relation to involvement and emotion instead of asking "were you engaged".

Additionally, as children (4-10 years old) may have communication issues, the Smiley-o-meter [73] was used to design the questionnaire for children. The Smiley-o-meter uses pictorial representations of emotional faces to depict the different level of satisfaction (based on a 1 to 5 Likert scale) as shown in Figure 2-4. It has been widely applied in many studies to measure interest and enjoyment as it is easy to complete and requires no writing by the children.



Figure 2-4. The Smiley-o-meter based on a 5-point scale [73].

2.5.2 External Observation

Researchers have argued that the self-report questionnaire may not be suitable for all users because the accuracy of answers relies on their interpretation of researchers' questions and the person's feelings at that time they filled out the questionnaire. A common strategy to overcome these relies on observations from external observers to measure engagement. For example, teachers may be asked to follow checklists to provide their subjective opinion of the extent to which their students are engaged. Human observers are commonly asked to follow checklists of measures that indicate engagement. To make rating more accurate, human observers may note examples of engaged behaviours on the score sheet during the observation.

Reference	Research area	Number of levels	Details
Bednarik <i>et al.</i> [10]	Conversational engagement	6	<ol style="list-style-type: none"> 1. No interest 2. Following 3. Responding 4. Conversing (without active discourse management) 5. Influencing discussion discourse/topic 6. Governing/managing discussion
Hernandez <i>et al.</i> [36]	TV viewers	4	<ol style="list-style-type: none"> 1. High 2. Medium 3. Low 4. None
Whitehill <i>et al.</i> [88]	Student engagement	4	<ol style="list-style-type: none"> 1. Not engaged at all 2. Nominally engaged 3. Engaged in task 4. Very engaged
Lee <i>et al.</i> [47]	Child education	4	<ol style="list-style-type: none"> 1. Interest high 2. Interest low 3. Boredom low 4. Boredom high

Table 2-5. Various levels of engagement annotation scales from previous studies.

There are two important aspects for designing a good rating scale for external observation: 1) choosing a proper number of levels; and 2) providing the specific evaluation criteria for each level about how the information will be labelled. In multiple engagement studies, researchers have defined various levels of engagement. For example, Bednarik *et al.* [10] designed an annotation scheme with six levels of conversational engagement ranging from “no interest” to “governing/managing discussion”. Table 2-5 shows several previous studies most of which used a scale with four categories [36, 47, 88]. Besides these studies, both child engagement measurement methods from Section 2.2.4 also had four levels: the E-Qual system [72] includes sophisticated, differentiated, focused, and unsophisticated; while the CEQ system [52] has a four-scale rating including: not at all typical, somewhat typical, typical, and very typical, where “typical” means that the child spends time in the activity. These studies show that engagement annotation scales with 4 levels can work for child engagement measurements in different contexts, which gives a support for the design of a general 4-level annotation scale to be used in this thesis.

Behaviours related to engagement are used to specify evaluation criteria of each level. For studies involving children, Hanna and colleagues suggested considering observed facial expressions of children, such as frowns and yawns, as a better engagement indicator rather than their answers to questionnaires [43]. Read *et al.* [73] measured child engagement as a useful dimension of ‘fun’. They observed and rated video recordings by a set of behaviours. Smiles were recognised as a positive behaviour while frowns and yawns were negative ones. Section 2.2.4 also provided several children’s engaged behaviours in a preschool classroom environment, such as watching or listening to the objects in the environment for a duration of at least 3 seconds.

Tests in this thesis were based on a child psychiatric study – Manchester Child Attachment Story Task (MCAST), introduced in Section 2.3. In MCAST assessment, child engagement is measured by a human assessor’s observation of a child’s facial expressions, using the MCAST protocol [32]. A rating scale of the extent to which the child engaged in the story relies on increasing attention to the play materials and the story, lack of distraction to other things, and the quality of emotional engagement in the story. Good engagement quality can be coded on behaviours such as an embarrassed laugh or smiling. The original engagement scale from the MCAST assessment protocol ranges from 1 to 9 and these 9 levels are grouped into four sections: normal/optimal range (score 7-9), borderline (score 5-6), abnormal scores (score 3-4) and seriously abnormal scores (score 3 and below).

An important aspect of external observation is to rate video recordings with a good timescale. Whitehill *et al.* [88] compared the effect of different timescales on human annotation in student engagement levels. They chose three timescales including: 1) watching video clips (without audio) and giving continuous engagement labels by pressing the Up/Down arrow keys; 2) watching video clips of 10s (without audio) and giving a single number to each the clip; 3) viewing static images and giving a single number to rate each image. They found approach 1) was hard to execute as it was difficult to provide accuracy labels for short moments as well as to provide continuous labels synchronised with the video clips. Compared to 1), approaches 2) and 3) were much easier to perform in the annotation task.

In summary, external observation is a common method for measuring engagement from different contexts and age groups. Human observers are usually asked to follow the checklist based on participants' facial expressions. This thesis will develop a 4-level scale for labelling both adults' and children's engagement levels, based on the MCAST assessment tool and other studies [36, 47, 52, 72, 88] which were rated using four different levels. It will be presented in the next chapter, to rate engagement from not engaged to fully engaged. Each level is operationally defined in accordance with specific facial behaviours, such as visual focus of attention, embarrassed laugh, and looking away from the screen. Meanwhile, approaches of video rating have been considered in this thesis. From previous studies, giving a single number to rate the video clips/static images will be used for constructing an automatic engagement classification.

2.5.3 Automated Measures

Although both self-reports and external observation with checklists are common and useful, they require a great deal of time and effort from researchers and observers. Objective measures can be recognised as a common strategy to overcome these drawbacks without recourse to direct questioning or human involvement [23]. Unlike questionnaire, each objective measure tends to target a very specific aspect of engagement instead of addressing a range of variables. One broad category of objective measures which has been used to infer engagement is physiological. Physiological data can be captured using a broad range of sensors and examples of sensors are: eye trackers, mouse pressure, biosensors (e.g. temperature, blood pressure, heart rate), and camera (e.g. face tracking, body posture). Studies in this thesis are based on the MCAST test and MCAST designers indicated that children's engagement may be disrupted by wearing the biosensors, such as ECG. Therefore,

another physiological measurement method, the focus of this thesis, is based on computer vision. Attempts using computer vision have been made to infer engagement from audio and video data by analysing cues from the face, body posture and hand gestures. The input facial signals are those such as head position and orientation [36], facial expression [31, 50, 88], eye gaze [12, 41, 54] and a combination of facial signals [36, 47, 94]. These kinds of information are relatively easy to collect in many situations as they can be captured by a broad range of accessible, inexpensive and un-intrusive sensors such as eye trackers and cameras. In the following section, two tools of automated engagement measurement methods based on computer vision will be introduced.

2.5.4 Automated Measure 1 – Eye-tracking Techniques

Eye-tracking is a technique for measuring the point of gaze (where one is looking at any given time) and the sequence in which the eyes shift from one location to another, recorded in real time [71]. It requires a device that can track the size of pupil and the location of the eye. One common eye-tracking device functions by shining an infrared light into the eye and capturing the light that passes through the pupil and is reflected back by the cornea to a video camera located on or near the screen [54]. The participant sits a known distance from a screen and a computer coordinates the position of the eyes and what appears on the screen. Participants' eye data can be collected while they are watching the screen without disrupting their attention.

Typical eye-movement metrics include fixations (pauses over informative regions of interest) and saccades (rapid movements between fixations). The fixation is the main measurement used in eye-tracking technology, which can reveal the amount of processing being applied to objects at the point-of-regard [71]. Saccades, which are quick eye movements occurring between fixations, are irrelevant for many studies as little or no visual processing can be achieved during a saccade [79]. It is only used to measure if readers are skipping letters in reading tasks. While a typical fixation duration varies between 50-600ms, the average duration of a saccade is 20-40ms. Due to the fast movement during a saccade, the image on the retina usually has poor quality. Therefore, effective information related to engagement was collected and analysed during the fixation period. Poole *et al.* [71] provided two analysis metrics taken from previous studies, one for fixation and another for saccade. There are five considerations for the fixation metric, as shown in Table 2-6, other fixation factors were based on specific contexts such as a text reading task.

Metric	Description
Numbers of fixations overall	The number of fixations is the fixation count in a given area. More overall fixations indicate less efficient search.
Fixation per area of interest (AOI)	More fixations on a particular area indicate that it is more noticeable, or more important, to the viewer than other areas.
Fixation duration	A longer fixation duration indicates difficulty in extracting information, or it means that the object is more engaging in some way.
Gaze (also referred to as “dwell, fixation cluster” and “fixation cycle”)	Gaze is usually the sum of all fixation durations within a prescribed area. It is best used to compare attention distributed between targets. It can also be used as a measure of anticipation in situation awareness if longer gazes fall on an area of interest before a possible event occurring.
On-target (all target fixations)	Fixations on-target divided by total number of fixations. A lower ratio indicates lower search efficiency.

Table 2-6. Five considerations of a fixation analysis metric [71].

Information gathered from eye-tracking technology has been used to help measure engagement in the field of user-system interactions [7, 10, 21, 26, 36, 42, 54, 59, 75, 81]. For instance, looking at the TV can be recognised as a good indicator of high engagement in a TV viewing task [36]. In the context of human-agent communication, researchers analysed fixations to measure the degree of engagement so that they can improve naturalness in human-agent communication according to the participant’s engagement behaviours [42, 59]. They demonstrated a model using the features of mutual gaze occurrence, gaze duration, and eye movement distance that could provide the best performance of the user’s conversational engagement estimation. There are also studies in relation to gaze behaviours between human and humanoid interfaces (i.e., robots) [75, 81]. Robots in these studies can recognise users’ engagement by mutual gaze information and mimic human gaze behaviours in conversations so that users could engage in mutual gazes with these robots, directing their gaze to them during the conversation.

However, there is not a standard set of gaze behaviours for engagement measurements. Eye data should be analysed according to specific contexts to make more accurate measurements. For example, teachers may think that a slow rate of reading can be a good indicator of engagement correlated with attention and comprehension in the context of self-paced reading but may find that an engaged student reads faster due to increasing interest [54]. Moreover, there are few studies related to the analysis of children’s gaze behaviours and

these studies are often focused on children with autism such as measuring their memory and emotional recognition [7, 26].

One focus of this thesis is to provide interpretations of gaze behaviours related to engagement in the context of MCAST story-stem viewing. Chapter 3 analyses gaze data from adults to compare the primary fixation measures across different engagement levels to human annotations. The analysis could provide a general set of gaze behaviours for measuring the engagement levels in MCAST story viewers. Chapter 4 uses the same analysis procedure for measuring children's engagement levels while watching the MCAST stories. If eye-tracking is a good indicator for automatic child engagement measurement, it would demonstrate that children's gaze behaviours contain information related to their engagement levels since previous studies are focused on children with autism.

2.5.5 Automated Measure 2 – Facial Expression Recognition

Besides eye-tracking, measuring facial expressions is another method based on computer vision for evaluating engagement. Hanna and colleagues suggested considering observed facial expressions of children as a better engagement indicator rather than their answers to questionnaires [34]. From the MCAST protocol, child engagement is measured by a trained assessor's observation of facial expressions, gestures, etc. [32]. A set of behaviours could be used to measure child engagement like smiling, laughing, concentration signs and excitable bouncing. Children's facial behaviours related to engagement need to be interpreted differently in different situations because engagement is a complex concept. For example, 'smile' is a good indicator of high engagement related to interest in a TV-viewing task while related to comprehension and attention in a learning task.

Methods to analyse facial cues from static images have been proposed for building automatic engagement recognition. The Facial Action Coding System (FACS) [25], developed by Ekman and Friesen, is a comprehensive method for objectively coding facial expressions in terms of individual facial Action Units (AUs), which uses the intensity of over 40 distinct facial muscles. For example, Whitehill *et al.* [88] compared FACS tools with other computer vision techniques to automatically detect university students' engagement from their facial expressions. They analysed the signals that human observers use to judge engagement from students' faces and automated the process using machine learning.

However, these methods were mainly trained on adults and there is little research related to the analysis of children's spontaneous facial expressions. One study [50] used facial Action Units to analyse changes in spontaneous facial expressions of children during a problem-solving task. It demonstrated that automated FACS coding can be applied to behavioural research to find differences in expressions between correct and incorrect educational trails from large amounts of video of spontaneous actions. Therefore, this thesis aims to detect if the FACS could be used to measure children's engagement levels in the context of story viewing.

The software employed in this thesis is OpenFace [9], a fully open source real-time facial behaviour analysis system including facial landmark detection, head pose as well as eye gaze estimation, and facial Action Unit recognition. It has been trained and tested on tens of thousands of manually coded images of adults' and children' faces from around the world. Facial gestures were coded in terms of manually annotated facial action units associated with upper face muscle movements around the eyes, eyebrows, and upper cheeks. For facial AU recognition, OpenFace uses Support Vector Machines (SVM) for AU occurrence detection and Support Vector Regression (SVR) for AU intensity detection [8]. It is able to recognise a subset of AUs, specifically: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45 as shown in Figure 2-5. AU28 is lip suck and AU45 is blink, which both cannot show in a static image.

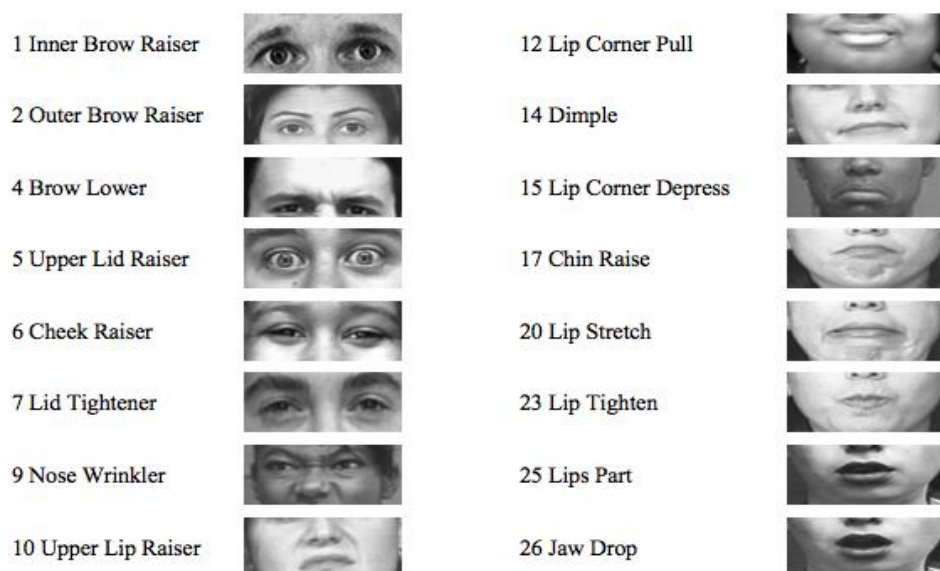


Figure 2-5. Sample facial action units from the FACS (Full AUs see⁹).

⁹ <https://imotions.com/blog/facial-action-coding-system/>

The facial action outputs will be used to measure children's engagement levels. If facial AU recognition is a good indicator for engagement measurement, it could be used for measuring children's engagement and can objectively capture the complexity of children's facial expressions. The time and involvement of MCAST assessors would be reduced. These findings will demonstrate if the facial data could be used across different age groups as a generalisable method for engagement measurement. If it is generalisable in the context of story viewing, researchers could collect large amounts of facial data to analyse the causes and variables that affect children's engagement in various applications to facilitate human-computer interaction, education and entertainment.

2.6 Conclusion

There are two key issues that Miller *et al.* [54] state are important for engagement research: 1) giving a definition of engagement based on the purpose of research; and 2) choosing the proper methods that match the definition. Since there is not a general definition of the term "engagement" in previous studies and it is interpreted based on different contexts and user groups of research, the definition of engagement used in this thesis was *a focusing of children's mood state around the particular distress represented in the story-stem*, where the context of this thesis is to digitalise the story-stem approach. The test to be used is an instance of the story-stem approach called MCAST, as discussed in Section 2.3. In order to reduce the time and cost required for MCAST administration and assessment, a system called the SAM is being developed to automate key steps of the MCAST test. The use of digital story-stems could reduce the time and effort of constructing the SAM system and improve the quality of data collected.

The second problem of engagement research is choosing the proper methods that match the definition. Section 2.5 introduced three methods to be used for measuring engagement: self-reports, external observation and automated measures. Automated measurement methods, the focus of this thesis, are based on computer vision, which provides an automatic estimation of engagement by analysing cues from the face. The input facial signals include eye gaze and facial expressions and can be collected without disrupting participants' attention while watching the digital stories. However, there is no previous work that measures child engagement using their eye gaze information. Also, studies for child engagement measurements using facial expressions were mainly focused on education, where the term engagement has a totally different interpretation than in psychiatric tests.

Engagement is positioned at children's mood state that focuses on attending to the play and materials, not being distracted by other things, and feeling empathy with the dolls and characters in the story. Since engagement in MCAST is measured by a trained assessor's observation of facial expressions, this thesis focuses on detecting if engagement could be reflected in modalities from the face, which aims to answer RQ1:

Can children's spontaneous facial expressions be used to automatically measure engagement levels in digital story-stems?

In addition, the questionnaire for children is more focused on children's attention and interest while they are asked to watch the stories to interpret their engagement levels. Children's answers to the questionnaire could be used as a support for investigating the accuracy automated measures. The combination of subjective and objective measures is a useful and efficient approach to gain a deeper understanding of child engagement. Meanwhile, automatic engagement recognition would reduce the time and involvement of MCAST administrators in running the tests. This will help assessors know whether the children are attending to a story and engaged in it; if they are not, then the test will not be successful. This is important for the SAM system as the aim there is to have SAM administrated by untrained users and for it to run automatically.

Research is on-going into automating the use of story-stems as the approach takes a lot of time and there are often few administrators trained to administer them. One motivation of this thesis is to create a more engaging digital story-stem used for the tests using the story-stem approach. Section 2.4 described different multimedia tools that could be used to create an engaging digital story-stem for children. The media factors to be investigated in the thesis are: animation vs. live-action video and storytelling voice due to easy implementation, which aims to answer RQ2:

How do voice type and presentation type affect child engagement levels in digital story-stems?

Chapter 3 An Initial Study of Adult Engagement Measurements

3.1 Introduction

The area of focus of this thesis is concerned with measuring children's engagement levels in digital story-stems. The story stems used are taken from the Manchester Child Attachment Story Task (MCAST), as introduced in Section 2.3. In MCAST, a deep engagement in a story-stem help children bring out their mental representation of attachment to their caregiver, which helps psychologists understand how children perceive the relationship with their caregivers (typically the parents) to be. According to this, engagement is defined in this thesis as *a focusing of children's mood state around the particular distress represented in the MCAST story-stem, which means that children focus on attending to the play and materials, are not distracted by other things, and feel empathy with the dolls in the story-stem.*

Since conducting MCAST assessments is expensive and time-consuming, a system called School Attachment Monitor (SAM) [78, 87] is being developed to automate the key steps of MCAST to reduce the time and cost required for MCAST administration and assessment. There are two novel aspects of SAM to improve the MCAST test: 1) automatically displaying the story-stem vignettes; 2) measuring children's engagement in the story-stems using their facial data. During the SAM test, children are asked to watch the story-stem vignettes shown on the screen then take over the controls of the PC and complete each story by speaking to the computer. Engagement in the stories is vital, as if the children are not engaged then poor-quality data will be collected and cannot be analysed to detect their Attachment status. For SAM, as in the original MCAST, engagement is measured when children are watching the vignettes to see whether the experience of 'distress' situation in the vignette will have activated their internal representation of attachment relationships and expectations of care.

In traditional MCAST assessment, engagement is rated by a trained assessor's observation of facial expressions from the video recordings, using the MCAST protocol. In SAM, the detailed facial movements of children when they are watching the vignettes can be captured in real time on a laptop. Using these facial movements, children's engagement levels can be measured automatically in each story-stem vignette. If a child is identified as 'disengaged' in one story-stem vignette, recordings of this child would not be used to assess the Attachment status based on their story and behaviour during the activity.

As discussed in Chapter 2, automated measurement methods based on computer vision have been used to indicate engagement from video recordings by analysing cues from the face. The input facial signals are those such as head position and orientation [36], facial expression [31, 50, 88], eye gaze [12, 41, 54] and a combination of facial signals [47, 93]. Section 2.5.4 discussed that eye-tracking technique has been used to measure engagement in previous studies. However, the eye-tracking measure was mainly trained on adults and there is little research related to the analysis of children's gaze behaviours. Studies related to children's gaze behaviours are focused on children with autism for measuring their memory and emotional recognition [7, 26]. Therefore, the first problem of this thesis is to investigate how to use the gaze data for measuring children's engagement levels while they are watching the digital story-stems.

Since the eye-tracking technique has been trained on adults, this chapter describes an initial experiment for evaluating adult engagement using their gaze behaviours in digital stories, which is *a preliminary study* for Chapter 4. This work was to develop and test the experimental framework needed for Chapter 4 and to provide foundations for RQ1. Two specific questions are asked:

Q1: What kinds of facial behaviours can be used for designing a coding system for the engagement levels in story-stem vignettes taken from MCAST?

Q2: What features of eye movement data should be analysed for different engagement levels?

3.2 Methods

3.2.1 Participants

Twenty university students (21-26 years old, 10 males and 10 females) were recruited from two UK universities participated in this study.

3.2.2 Procedure

The test took approximately 20 minutes for each adult. To start, an introduction to the procedure was given and the participant's eye movements were calibrated using the Tobii's calibration procedure¹⁰. Then participants were asked to watch the four 'distress' MCAST story videos that were presented on a screen. The full scripts were shown in Section 2.3.

During the watching session, the adult's facial behaviours were recorded by a Logitech C920¹¹ webcam and eye gaze data were collected using a Tobii EyeX eye-tracker¹². The eye-tracker controller collects the adult's gaze points on the screen and transforms them to pixel coordinates on the screen. The eye-tracker was placed at bottom of a laptop screen and the laptop was put in front of the adult. To establish temporal correspondence between the two recording systems (i.e. the eye-tracking signals and webcam video), a video play button was used and displayed on the touchable laptop screen. When the participant was ready to watch the story-stem, he/she was asked to press the play button and the story-stem started playing, which was recorded as a timestamp of the start of the trial. As the length of each story-stem was known (see Figure 3-1), there is no need of a signal used for the end of the trial. The eye-tracking were synchronized using the start timestamp to obtain accurate temporal correspondence.

3.2.3 Data Annotation

The timescale for annotating video recordings at which labelling takes place: clips of 10 sec

Once the videos had been recorded for all adults, the next step was to label them in terms of engagement. Section 2.5.2 discussed one method of annotation was viewing the recordings as short video clips of 10s (without audio) and giving a single number to rate each clip. Thus,

¹⁰ <https://help.tobii.com/hc/en-us/articles/209530409-Create-a-new-user-profile>

¹¹ <https://www.logitech.com/en-gb/product/hd-pro-webcam-c920>

¹² <https://developer.tobii.com/an-introduction-to-the-tobii-eyex-sdk/>

all video recordings were split into 10-second clips and each clip was given a number to label the engagement level. Each adult had 4 recordings for 4 story-stems and the length of each recording was 27s, 32s, 36s, and 85s, the same as the length of each story-stem. Recordings of each story-stem cannot be divided evenly; for example, recordings of the ‘nightmare’ vignette would be split into two 10-second clips and one 7-second segment that cannot be used. To ensure that all recordings would be used, the four recordings of each adult were integrated into one video (180s in total for each integrated recording) and then split into 10-second video clips, as shown in Figure 3-1. Therefore, the length of recording for each adult was 180s that can be split into 18 clips (S1~S18), 20 adults had 360 clips in total.

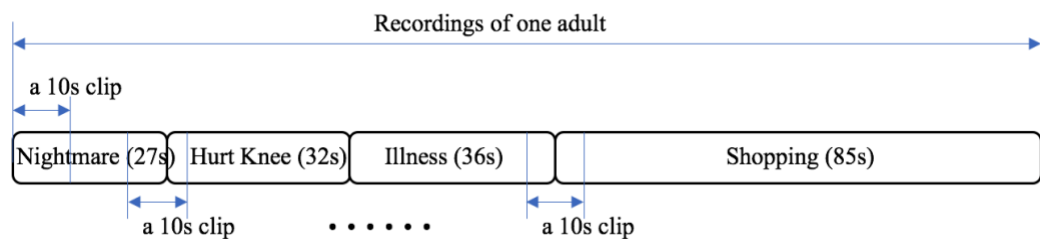


Figure 3-1. The way of splitting the video recording of one participant.

The next step was to describe the annotation process in relation to 1) the people who did the annotations; 2) the way of presenting the data to human labellers and 3) the recordings of their annotation results for each video clip.

Details of human labellers and the annotation process

Five labellers (L1~L5) were used to annotate the clips in terms of engagement. They were PhD students studying computing science and mathematics. One of them (the author of the thesis, L1) has attended the official MCAST training and has certification to administer the MCAST test. The others (L2~L5) had a brief MCAST training, including an explanation of MCAST administration and assessment based on the MCAST papers [32], the engagement rating scale taken from the MCAST training documents, and a practice of rating several MCAST videos (already permitted and rated by MCAST experts).

Based on the definition, the engagement level is a rating of representing the extent to which the child is absorbed and imaginatively caught up in each MCAST story-stem. It is rated based on cues such as: increasing attention to the story, lack of distraction to other things, and feeling empathy to the doll on the screen, as measured by their facial expressions.

After the brief MCAST training, labellers were asked to independently view and rate each clip based only on the participants' facial appearance. All clips were randomly allocated to the labellers and, when labelling video clips, the audio was turned off so that labelling was based only on facial expressions. There were 360 clips in total and each clip was annotated by two labellers. Thus, each labeller was allocated to rate 144 clips and they were asked to finish the rating in 5 days.

In order to design an annotation scale for measuring adult engagement, previous studies showed that engagement annotation scales with 4 levels can work for both child and adult engagement measurements in different contexts (see Table 2-5). Thus, the five labellers were asked to annotate each clip with an engagement level $l \in \{1, 2, 3, 4\}$ (1 = not engaged, 2 = rarely engaged, 3 = highly engaged, and 4 = fully engaged). Meanwhile, they were also asked to note examples of engaged behaviours on the rating form (see Figure 3-2) during the observation. The rating form was designed using a spreadsheet that contained two columns to record the engagement rating and notes of engaged/disengaged behaviours respectively. If one clip was unclear (e.g. no eyes, eye/face occlusion) or contained no person at all, they were asked to annotate this clip with an X. The instruction sheet to annotators is shown in Appendix G.

Labeller:		
Allocated clips	Engagement level	examples of engaged/disengaged behaviours
1		
2		
3		
4		
5		

Figure 3-2. The rating form used for human labellers to record the engagement rating and engaged/disengaged behaviours.

After the data annotation, there were 19 clips labelled as X with agreement from all labellers due to bad quality and eye/face occlusion. The next step was to 1) calculate the agreement of engagement ratings between the two independent labellers and 2) summarise labeller' notes of examples of engaged/disengaged behaviours. Both steps aimed to generate a scale of the engagement level annotation categories.

Annotation agreement and examples of engaged behaviours from human labellers

Firstly, one participant's rating scores of engagement were shown as an example of rating results about how engagement varied across different clips. P1 indicates this data was taken from a participant that has an ID as P1. As discussed earlier, the recording for each participant can be split into 18 clips (S1~S18), so there were 18 clips (P1S1~P1S18) shown in Figure 3-3. S1 and S2 are from the first story-stem; S3 merges the end of the first story-stem and the beginning of the second story-stem (70% for story-stem 1 and 30% for story-stem 2); S4 and S5 are from the second story-stem; S6 merges the end of the second story-stem and the beginning of the third story-stem (90% for story-stem 2 and 10% for story-stem 3); S7-S9 are from the third story-stem; S10 merges the end of the third story-stem and the beginning of the fourth story-stem (50% for story-stem 3 and 50% for story-stem 4); S11-S18 are from the fourth story-stem.

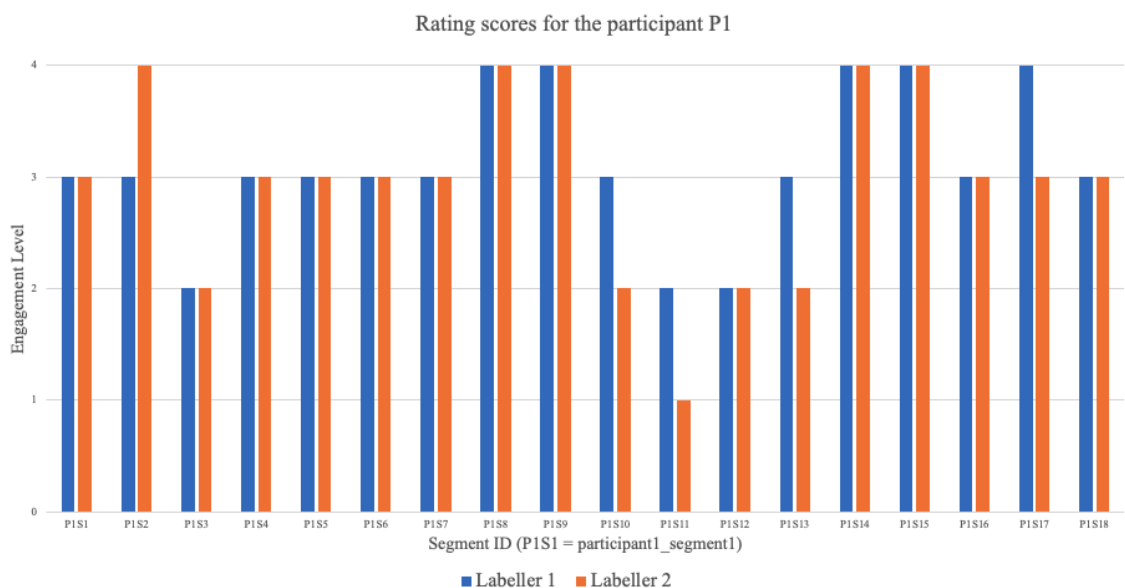


Figure 3-3. The rating scores of participant P1's engagement levels watching the four MCAST story-stems from two independent labellers.

Figure 3-3 shows that this participant was engaged during much of the time watching the story-stems. The agreement of rating scores was 72.22% (13 clips with consistent ratings over 18 clips). Although S3, S6 and S10 are segments that merged together the clips, the participant's engagement levels did not vary significantly (e.g., S6 has the same scores compared to S5 and S7.). Also, the agreement of the 'merging' clips was 66.67% (S3 and S6 have a consistent rating score while S10 has not.) Therefore, merging together the clips in this way is an available way to collect more data for measuring the engagement levels.

Then the inter-rater reliability, the degree of agreement between the two raters, was performed using a weighted Cohen's κ ^{13,14}. The weighted kappa is calculated using a pre-defined table of weights (see Table 3-1(right)) which measures the degree of disagreement between the two labellers, the higher the disagreement the higher the weight.

The distribution of annotation scores is shown in Table 3-1 (left). Cohen's κ was 0.825 (s.e. = 0.023) in this dataset, which means that labellers have an almost perfect agreement (0.81~1.00) in rating the engagement levels. The percentage of agreement, the proportion of clips with the two same scores, was $287/341 = 84.16\%$. There were 190 clips (55.72%) that were labelled as 'engaged', including level 3 and 4. Cohen's κ was 0.748 (s.e. = 0.052) in this 'engaged' dataset, which means that labellers have a substantial agreement (0.61~0.80) in labelling the engaged data. The proportion of clips which the two scores were the same was 88.95% over all engaged clips.

		Labeller 1 Engagement Score				
		1	2	3	4	Total
Labeller 2 Engagement Score	1	38	5	3	0	46
	2	7	80	8	1	96
	3	2	6	118	12	138
	4	0	1	9	51	61
	Total	47	92	145	64	341
Agreement		38	80	118	51	287

0	1	2	3
1	0	1	2
2	1	0	1
3	2	1	0

The table of weights. The grey table was the weights of engaged data.

Table 3-1. (left) Overall engagement ratings (19 clips labelled as X due to quality). (right) The table of weights. The agreement across engagement levels 1, 2, 3, and 4: 84.16%: Cohen's kappa = 0.825. Engaged data (level 3 and 4) agreement 88.95%: Cohen's kappa = 0.748.

The examples of engaged/disengaged behaviours taken from the rating forms showed that a set of facial behaviours including gazes, smiles, frowns, and eyebrow movements was identified. For example, looking away from the screen was recorded as a disengaged behaviour when the clip was labelled as level 1. A summary of engagement ratings and

¹³ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/>

¹⁴ <http://www.real-statistics.com/reliability/interrater-reliability/weighted-cohens-kappa/>

examples of engaged/disengaged behaviours on the rating form for the participant (P1) is shown in Figure 3-4.

Participant: P1						
Clips	Labeller 1			Labeller 2		
	labeller ID	Engagement level	examples of engaged/disengaged behaviours	labeller ID	Engagement level	examples of engaged/disengaged behaviours
S1	L3	3		L4	3	
S2	L2	3		L3	4	watching carefully
S3	L4	2		L2	2	
S4	L5	3		L1	3	
S5	L1	3	looking at the screen to watch the story	L5	3	
S6	L1	3	eye movement	L3	3	watching carefully
S7	L5	3		L4	3	
S8	L3	4	brow movement	L1	4	looks surprised, mouth open, fully engaged
S9	L2	4	fully engaged, behaviours include brow, eye and mouth	L4	4	
S10	L5	3		L3	2	
S11	L4	2		L1	1	looking away for a while + lip corner down
S12	L4	2		L2	2	Watching the bottom of the screen or the keyboard of the laptop
S13	L3	3		L5	2	
S14	L2	4	level 4, big eyes staring at the screen	L1	4	
S15	L5	4		L2	4	eye movement, seems watching the movement of characters on the screen
S16	L1	3		L4	3	
S17	L4	4		L3	3	
S18	L3	3		L5	3	

Figure 3-4. Annotation results for the participant P1 taken from rating forms. Clips (S1~S18) were randomly annotated to five labellers (ID: L1~L5). The blank areas mean that no engaged/disengaged behaviours were recorded by labellers.

Generation of the engagement level annotation categories

The engagement annotation categories were shown in Table 3-2. In this table, the level and name were taken from the given engagement level $l \in \{1, 2, 3, 4\}$ to labellers (1 = not engaged, 2 = rarely engaged, 3 = highly engaged, and 4 = fully engaged). The characteristic of each level was described using recorded engaged/disengaged behaviours from the labellers. The performance of this scale will be tested in the next chapter for annotating children's engagement levels so that it could be generalised across different age groups.

Level	Name	Characteristic
1	Not engaged	e.g. looking away from screen and focusing on something other than the story; eyes completely closed over 3 seconds.
2	Rarely engaged	e.g. clearly not "into" the story; paying attention to something else (e.g. camera and desktop eye-tracker), but sometimes focusing on the story
3	Highly engaged	e.g. good enough to proceed to the task; participant requires no admonition to "stay on task"
4	Fully engaged	e.g. good quality engagement; participant could be "commended" for his/her level of engagement in task
X		The clip was very unclear or contains no person at all.

Table 3-2. The engagement level annotation categories used by the labellers.

3.2.4 Data Selection

After data annotation, the next stage is data selection to ensure that each clip has a rating score of the engagement level. Since the length of the recording for each adult was 180s that can be split into 18 ten second clips, 20 adults totally had 360 clips that annotated by the 5 labellers and each clip was independently annotated by 2 labellers. The annotation stage indicated that there were 19 clips labelled as X, with agreement from all labellers, due to bad quality and eye/face occlusion. There were 341 clips annotated by human labellers and 287 clips had two scores the same (see Table 3-1). The following procedure was used to select data for analysis and classification:

- 1) If either labeller marked a clip as X (no eyes, eye occlusion, or unclear), the clip was discarded;
- 2) If two scores for a clip were the same, it was retained;
- 3) If the two scores were not the same, (e.g., one labeller assigns a label of 1 and another assigns a label of 2/3/4), the other three labellers would label this clip and the labelling result was the average labelling across all five labellers. The “ground truth” label for the clip was computed by rounding the average label for that clip to the nearest integer (e.g., 2.4 rounds to 2; 2.5 rounds to 3).

For example, the rating scores of 18 clips for the participant (P1) are shown in Figure 3-3. Each clip was annotated by two independent labellers. There were 5 clips (S2, S10, S11, S13, and S17) with different engagement ratings from the two labellers. According to the step 3) of data selection procedure, these 5 clips were annotated by the other three labellers (e.g., S2 was rated again by L1, L4 and L5 as the rating score from L2 and L3 was not consistent) and give an average rating across all five labellers. After the data selection procedure, the final engagement rating result of 18 clips for participant P1 is shown in Table 3-3.

Clips	S1	S2	S3	S4	S5	S6	S7	S8	S9
Engagement level	3	3	2	3	3	3	3	4	4
Clips	S10	S11	S12	S13	S14	S15	S16	S17	S18
Engagement level	3	1	2	2	4	4	3	3	3

Table 3-3. The final rating result of engagement level $l \in \{1, 2, 3, 4\}$ of 18 clips for participant P1 using the data selection procedure.

In total, there were 341 clips selected using this approach. The distribution of clips in each level of engagement is shown in Table 3-4 from level 1 to level 4 respectively.

Level of Engagement	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Count of clips (%)	46 (13.49%)	93 (27.27%)	137 (40.18%)	65 (19.06%)

Table 3-4. The distribution of adult engagement levels, shown in the count of clips and in parenthesis in percent.

3.3 Eye-tracking measures

Once the data had been annotated, there are two steps to analyse the relationship between the engagement levels and gaze behaviours. Since fixations are the most common feature of eye-tracking studies, recognising fixations was the first important problem for this test. The first step was to analyse fixations by comparing the primary fixation metrics among different levels of engagement. Although the Tobii EyeX Engine calculates fixations in real-time, the real-time filter cannot provide a stable data stream of eye coordinates (Tobii would not recommend using a real-time fixation filter in academic research, see the Tobii’s FAQ website¹⁵). Therefore, gaze data will be collected by using the Tobii EyeX eye-tracker and these gaze data will be grouped into fixations using the following algorithms. The algorithm of fixation identification and the default values of determining the fixation filter was taken from Tobii’s public documents [66]. The second step was to develop a classifier that could identify adult engagement levels using the fixation metrics.

The following three sections discuss: 1) how gaze data in each clip was grouped into fixations using the I-VT algorithm; 2) what fixation metrics are computed; and 3) how to conduct a classification task to identify adult engagement levels using the fixation metrics.

3.3.1 Fixation Identification

Gaze¹⁶ is the raw data collected in this test for constructing fixations, which is the spatial locations of the visual landings on the stimulus instantaneously¹⁷. It can be collected using eye trackers. Eye-tracking in this thesis was performed using a Tobii EyeX eye-tracker. It is a desktop eye-tracker, which can be added to a regular computer screen and will not disrupt

¹⁵ <https://developer.tobii.com/community/forums/topic/algorithm-used-in-fixationdatastream/>

¹⁶ To avoid confusion, the term gaze in this thesis means the raw data. The factor “gaze” of the fixation metric in Table 2-6 will be described as “fixation cluster”.

¹⁷ <https://www.tobii.com/learn-and-support/learn/steps-in-an-eye-tracking-study/data/how-are-fixations-defined-when-analyzing-eye-tracking-data/>

participants' while watching the digital stories. During a recording, the Tobii EyeX eye-tracker collects raw eye movement data points at a sampling rate of 60Hz, which means a gaze point will be recorded every 16.7 milliseconds. Each gaze data point is identified by a timestamp and (x, y) coordinates. The coordinate system used in this thesis was the screen coordinate system in relation to the pixels on the screen as digital stories were displayed on the full screen. Figure 3-5 shows an example of one participant's gaze data of watching the 'Hopscotch' story-stem and also a screenshot of this story-stem.

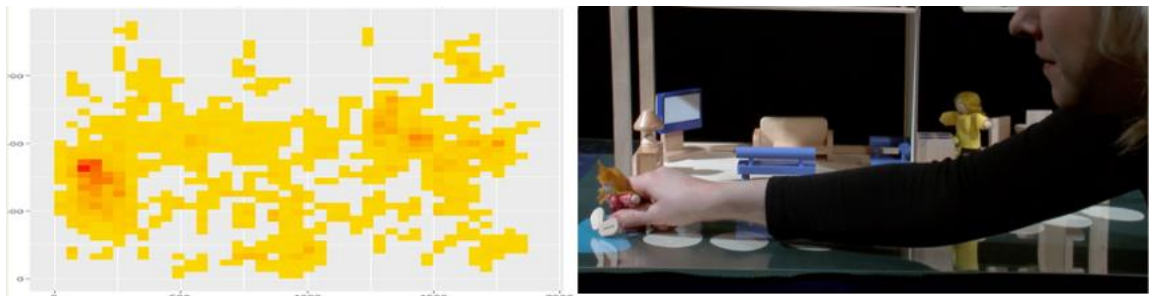


Figure 3-5. The heatmap of one participant's gaze data and a screenshot of the corresponding story-stem.

The heatmap of the gaze data showed that more gaze data were located as the position of the child doll (red dress) on screen. The next step was to group the gaze data into fixations. Firstly, to describe the engagement level and gaze patterns, a large number of gaze data were constructed using a fixed 'window length'. The parameter 'window length' is used to specify the length of the fixed period. In this research, during each 10 second the raw gaze data were computed and already labelled according to the corresponding engagement level from human annotations.

In order to identify fixations within the raw gaze data using a fixed window size interval, there are various types of fixation identification. For example, Tobii provided three filters including the Tobii Fixation, the ClearView Fixation Filter and the Velocity-Threshold fixation Identification (I-VT) [66]. Among these fixation classification algorithms, the I-VT classification algorithm is the most common as it is relatively easy to implement and to understand [66]. The Velocity-Threshold Identification (I-VT) is a velocity-based filter to calculate if a point-to-point velocity has a lower or higher speed than a pre-defined threshold. The velocity is most commonly given in visual angle, i.e., degrees per second ($^{\circ}/s$). It is computed based on the distance between the eye position and screen as well as the gaze positions on the screen. The angle velocity is calculated between the eye position and two gaze points, then divided by time between two gaze samples [79]. The average angle velocity

would solve the problem that only one gaze point with the given clips as the velocity calculation needs at least two samples.

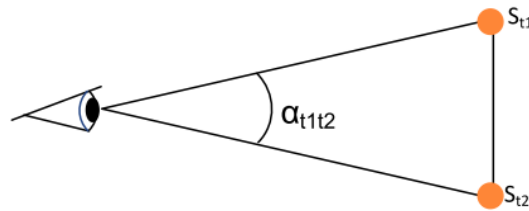


Figure 3-6. Calculation of gaze angle [66].

If this sample, the angle velocity associated with two consecutive gaze points, is below the ‘velocity threshold’ parameter, that sample is classified as belonging to a fixation; and if it is the same or higher than the parameter, it is classified as a saccade. Once the sample has been classified, it compares to the preceding sample. If the current and the preceding sample have the same classification, the current sample is added to the list that includes the previous sample as well as all consecutive samples with the same classification. If the current sample and the preceding sample have different classifications, the current sample will be added to a new list and the next sample would be classified.

If fixations are located close together both in time and positions on a screen, it is very likely that they are parts of one long fixation. Since the fixation duration is an important factor of the fixation metric, the next step is to merge adjacent fixations to avoid a long fixation to be divided into several short ones due to very short saccades. To test if two consecutive fixations should be merged, the first thing that is to specify a maximum time between two fixations. The default value is 75ms taken from Tobii’s report [66] and it has been validated in previous studies. Then the time between the end of the first fixation and the beginning of the next is compared to the default value. If the time is shorter than the specified maximum time, the two fixations are merged. Moreover, a very short fixation cannot be used for analysing human behaviours as the eye and brain need some time for receiving the incoming information [57]. These short fixations should be discarded by comparing it to a specific minimum fixation duration value. The default value of the minimum fixation duration is 60ms [66]. If the duration is shorter than the parameter value, this fixation will be discarded or reclassified.

Therefore, there are three steps in this test for grouping adults’ raw gaze data into fixations including classifying the gaze sample into fixations, merging adjacent fixations and

discarding short fixations. The parameters of eye-tracker used in this test includes the sampling frequency (60Hz), display resolution for identifying gaze point (1920x1080) and a viewing distance of 50cm. In this test, adult's raw gaze data were grouped into fixations using the Velocity-Threshold Identification (I-VT) algorithm. The best value of the velocity threshold parameter in the I-VT classifier for identifying stable fixations is $30^{\circ}/s$ [65] and it will be used in this thesis as it is the most suitable value for an eye-tracker with a sampling frequency of 60Hz. The first gaze point was added a label "start" and computed the angle velocity between this point and the next gaze point. If this velocity was below the threshold, the next gaze point was labelled as "data", which means that the two gaze points belongs to one fixation. Then the next gaze point was the current point and the velocity was computed between the current point and the next point. While the velocity was higher than the threshold, the next point was labelled as "end". Then the point after "the next point" would be the current point that labelled as "start". After this step, a fixation was from a gaze point labelled as "start" to the nearest next gaze point labelled as "end". The second and third step was to merge adjacent fixations and discard short fixations. The maximum time between fixations was 75ms used for merging adjacent fixations and the minimum fixation duration was 60ms used for discarding short fixations¹⁸. For merging the adjacent fixations, the duration was computed between the current fixation and the next fixation. If the duration was shorter than the maximum time, the two fixations are merged. For discarding short fixations, the duration was computed from the timestamp of the "start" point to the timestamp of the "end" point for each fixation. If the duration was shorter than the minimum fixation duration, the fixation was discarded. Since fixations have been classified, the next step is to detect whether fixations contain information related to adult engagement by comparing the fixation metrics.

3.3.2 Fixation metrics

There are five considerations for a fixation metrics as shown in Table 2-6 including: the overall number of fixations, fixation duration, fixation per area of interest (AOI), fixation cluster and on-target fixations. Fixation per area of interest (AOI), fixation clusters and on-target fixations aim to measure fixations in a prescribed or a specific target area, however there is no prescribed area or specific target area when watching the digital story-stem vignettes. Therefore, the following two fixation metrics were computed from the grouped

¹⁸ These default values under ideal conditions were taken from [66].

raw gaze signals using fixed window size approach and in terms of adult engagement levels: the overall number of fixations and fixation duration.

The overall number of fixations. The overall number of fixations is the total number of fixations in a given area. A higher overall fixation number indicates that the area is more noticeable or important to the participant than other areas. This feature was computed by the number of fixations across the different engagement levels.

Fixation duration. Fixation duration is the length of a fixation. Longer fixation durations mean higher comprehension or more difficulty in extracting information from that area. The higher comprehension means that a slow rate of reading can be a good indicator of engagement correlated with higher comprehension in the context of self-paced reading. A corollary in this thesis was participant's gaze behaviour changes from a slow-paced information extraction to a higher comprehension in the context of story video viewing. This feature was computed by the fixation duration including the mean fixation duration and the sum of fixation durations across the different engagement levels.

3.3.3 Classification

Recordings were annotated by human labellers with four classes. As the classifier aims to detect whether the adult was engaged or disengaged in the digital story-stems, a 4-class classification task was turned into a classification problem with two classes: low engagement levels and high engagement levels. Class A was the low engagement levels including the not engaged and the rarely engaged categories while the highly engaged and fully engaged categories were grouped into class B for high engagement levels. To perform the classification task, the LIBSVM [19] was used as an efficient implementation of the standard soft-margin Support Vector Machine (SVM) [20]. Each data was normalised by linear mapping into an interval [0,1]. Second, the SVM classifier was created to detect whether an adult is engaged in the digital story-stem vignettes from the extracted fixation durations. In designing classifiers, the dataset is split into training (60 to 80%) and testing (40 to 20%) in the literature because the 60 to 80% for training is to better model for the underlying distribution and then test the results with the remaining 40-20%. In this study, the 70-30 ratio was chosen as this ratio was the average value of previous classifiers, which means that recordings with valid gaze data was divided into two sets: one with 70 percent of the source data, for training the model, and one with 30 percent of the source data, for testing the model.

3.4 Results

There are two fixation metrics to be used: the overall number of fixations and fixation duration. A statistical analysis is presented about the fixation metrics with regard to the different levels of adult engagement based on results from human annotation and a classification using adults' fixation metrics is then performed where each clip is marked as being either high or low engagement.

3.4.1 Two Fixation Metrics

To detect if adults' gaze behaviours contain information related to their engagement levels during watching the story-stem videos, this section presents a statistical analysis of the two fixation metrics with regard to the different levels of adult engagement.

Twenty adult participants (recorded as P1~P20) performed a total of 80 story-stem trials (4 story-stems for each participant). The video recordings of 4 story-stems for each adult were integrated into one video recording firstly (180s in total for each integrated video recording) and split into 10-second video clips (18 clips per adult, recorded as S1~S18). The raw data were collected by an eye tracker and each gaze data point was identified by a timestamp and (x, y) coordinates (see examples in Figure 3-7 (left), taken from a 10-second clip with ID: P1S1). The coordinate system used here was the screen coordinate system in relation to the pixels on the screen as digital story-stem videos were displayed on the full screen. The timestamp reference point in microseconds was saved as an arbitrary point in time.

Aggregating data was performed to group the raw gaze data into fixations, as introduced in Section 3.3.1. Since there is no way to set a custom timestamp¹⁹, the way to calculate the duration of each fixation was to save the first received timestamp and the timestamp for the next gaze point, then subtract the first gaze data timestamp from the second and compute how many microseconds have passed, and then convert the microseconds into milliseconds. Figure 3-7 (right) shows the duration of each identified fixations in each clip. For example, the row with clip ID P1S1 included 14 fixations and the duration of the first fixation was 368ms. Overall, the 341 clips (clip ID: P1S1~P20S18, 19 discarded due to eye/face occlusion) contained 4228 fixations.

¹⁹ No document to exemplify the usage of timestamp. The timestamp was only used to calculate the duration (<https://developer.tobii.com/community/forums/topic/acquiring-current-timestamp/>).

clip ID	Gaze Data
P1S1	Gaze Data: (922.4, 707.6) timestamp 18608110 ms
P1S1	Gaze Data: (923.9, 706.0) timestamp 18608124 ms
P1S1	Gaze Data: (929.2, 709.1) timestamp 18608140 ms
P1S1	Gaze Data: (928.3, 704.7) timestamp 18608155 ms
P1S1	Gaze Data: (927.6, 703.9) timestamp 18608169 ms
P1S1	Gaze Data: (928.6, 704.7) timestamp 18608201 ms
P1S1	Gaze Data: (929.8, 701.7) timestamp 18608215 ms
P1S1	Gaze Data: (932.0, 695.6) timestamp 18608230 ms
P1S1	Gaze Data: (932.4, 691.9) timestamp 18608245 ms
P1S1	Gaze Data: (933.3, 685.9) timestamp 18608259 ms
P1S1	Gaze Data: (931.1, 680.4) timestamp 18608275 ms
P1S1	Gaze Data: (934.7, 683.6) timestamp 18608290 ms
P1S1	Gaze Data: (936.0, 692.5) timestamp 18608304 ms
P1S1	Gaze Data: (940.6, 694.4) timestamp 18608320 ms
P1S1	Gaze Data: (944.2, 691.4) timestamp 18608350 ms

Clip ID	Engagement	Fixations (recorded by duration(ms))													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
P1S1	3	368	435	336	517	283	243	492	124	286	325	132	122	695	269
P1S2	3	551	236	390	367	195	388	60	449	185					
P1S3	2	216	162	147	284	93	290	281	372						
P1S4	3	117	389	191	634	487	354	566	347						
P1S5	3	457	542	426	128	417	341	436	420						
P1S6	3	219	258	542	617	586	294	533	294	68					
P1S7	3	388	621	575	477	433	506	518	357	579	524				
P1S8	4	301	330	539	276	720	777	1187							

Figure 3-7. The screenshots of collected and aggregated data. (left) Raw gaze data was collected by the eye tracker. (right) The raw gaze data were grouped into fixations and combined with the engagement ratings from human annotation.

For the column titled ‘engagement’ in Figure 3-7 (right), the level of engagement in each clip was taken from the annotation results of video recordings for all adults. The number (percentage) of the selected 341 video clips (not appropriately recorded for 19 clips) in each level of engagement was shown in Table 3-4. It shows that there were some relatively rare classes, such as not engaged and fully engaged of the story videos viewing. Therefore, the aggregated data included three categories: clip ID, engagement level and fixations (recorded by durations), shown in Figure 3-7 (right).

The overall number of fixations

In this section, analysing the overall number of fixations would determine whether this fixation metric contained information related to engagement level during watching the story-stem videos. The variables used here were computed as a set of the-number-of-fixations/the-level-of-engagement pairs in each 10-second clip, which were recorded in the forms of numbers (see Figure 3-8 (right)).

Clip ID	Engagement	Fixations (recorded by duration(ms))														Number of Fixations
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	
P1S1	3	368	435	336	517	283	243	492	124	286	325	132	122	695	269	14
P1S2	3	551	236	390	367	195	388	60	449	185						9
P1S3	2	216	162	147	284	93	290	281	372							8
P1S4	3	117	389	191	634	487	354	566	347							8
P1S5	3	457	542	426	128	417	341	436	420							8
P1S6	3	219	258	542	617	586	294	533	294	68						9
P1S7	3	388	621	575	477	433	506	518	357	579	524					10
P1S8	4	301	330	539	276	720	777	1187								7
P1S9	4															5
P1S10	3															21
P1S11	1															12
P1S12	2															5
P1S13	2															8
P1S14	4															7
P1S15	4															8
P1S16	3															9
P1S17	3															11
P1S18	3															10

Figure 3-8. The number of fixations per clip was computed from the aggregated data. (left) Same as Figure 3-7 (right) to show the duration of each fixation in one 10s clip. (right) The pair of data (engagement, the number of fixations) was used in the ANOVA test.

In each pair of data, the level of engagement $l \in \{1, 2, 3, 4\}$ was a rating score taken from human annotation and selection results; fixations were identified within the raw gaze data and recorded by its duration (see Figure 3-8 (left), same as Figure 3-7 (right)) and the number of fixations was computed per clip. For example, Figure 3-8 (left) shows that the first row (clip ID: P1S1) included 14 elements in the ‘fixation’ category. This row was converted into the first row (the same clip ID: P1S1) in the right graph to be analysed.

Firstly, a descriptive statistical analysis shows the total number of fixations and the average number of fixations per clip according to four engagement levels. Overall, the 341 clips contained 4228 fixations and participants averagely would make about 12 fixations during a 10-second clip. The total number of fixations increased when the engagement levels increased from level 1 to 3, while it decreased during level 3 to 4 (see Table 3-5). For example, the highly-engaged category (level = 3) has the highest total number of fixations because clips labelled as highly-engaged are the most frequently occurring (40.49%). The average number of fixations per clip (= the total number of fixations / clip counts) was then computed for engagement level 1 to 4 (see Figure 3-9). The average number of fixations per clip decreased when the level of engagement increased but there was not much difference between level 1 and 2.

Level of Engagement	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Clip Count (%)	46 (13.49%)	93 (27.27%)	137 (40.18%)	65 (19.06%)
Total number of Fixations (%)	636 (15.04%)	1248 (29.52%)	1712 (40.49%)	632 (14.95%)

Table 3-5. The overall distribution of adult engagement and the total number of fixations, shown in counts and in parenthesis in percentages.

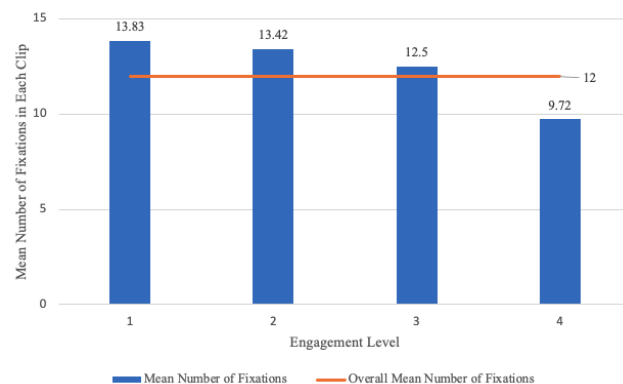


Figure 3-9. The overall distribution of the engagement levels and the average number of fixations per clip. The line shows the overall mean number of fixations of all fixations.

A one-way ANOVA was then performed to test for an effect of engagement level on the number of fixations per clip. The test modelled the differences in the mean of the response variable, the overall number of fixations per clip, as a function of the level of engagement. It indicated that there was a statistically significant difference in the overall number of fixations per clip according to the levels of engagement ($F(3, 337) = 4.994$, $p = .002$ under a significant level 0.05). To investigate where the actual differences were in the ANOVA test, a Hochberg²⁰ *post hoc* test was conducted as pairwise comparisons among groups for the independent variable (engagement level). Table 3-6 shows the differences in the mean number of fixations between engagement levels, the p-value and its standard error for multiple pairwise comparisons.

Dependent Variable: The number of fixations in each clip				
Engagement Level l	Engagement Level l'	Mean Difference ($l - l'$)	Std. Error	p-value
1	2	.407	1.194	1.000
	3	1.330	1.129	.805
	4	4.103*	1.276	.009
2	1	-.407	1.194	1.000
	3	.923	.890	.881
	4	3.696*	1.071	.004
3	1	-1.330	1.129	.805
	2	-.923	.890	.881
	4	2.773*	.998	.034
4	1	-4.103*	1.276	.009
	2	-3.696*	1.071	.004
	3	-2.773*	.998	.034

Table 3-6. Results of a Hochberg *post hoc* test to find the actual differences of the four engagement levels on the overall number of fixations per clip. *: The mean difference is significant at the 0.05 level.

The results of the *post hoc* tests show significant pairwise mean differences in the number of fixations per clip between the fully-engaged category (level 4) and the other three engagement levels separately. There was an average difference of 4.103 ($p = .009$) between level 1 and 4; an average difference 3.696 ($p = .004$) between level 2 and 4; and an average difference of 2.773 ($p = .034$) between 3 and 4. However, the differences in the number of fixations between other engagement levels, such as engagement level 1~2, 1~3 and 2~3,

²⁰ A Hochberg's post hoc test: Multiple comparison and range test that uses the Studentized maximum modulus. Similar to Tukey's honestly significant difference test. (see https://www.ibm.com/support/knowledgecenter/de/SSLVMB_23.0.0/spss/base/idh_ones_post.html)

were not statistically significant. Therefore, the statistical analysis shows that the overall number of fixations per clip only had a significant difference between the fully-engaged category and other three engagement levels respectively and is also not a good indicator for the engagement classification in the next section.

Fixation duration

Fixation duration is the length of a fixation to be used to analyse whether it contained information related to adult engagement levels during watching the story-stem videos. In this section, the variables was a set of fixation-duration/the-level-of-engagement pairs, which each pair of data was recorded in the forms of numbers (see Figure 3-10 (right)).

Clip ID	Engagement	Fixations (recorded by duration(ms))													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
P1S1	3	368	435	336	517	283	243	492	124	286	325	132	122	695	269
P1S2	3	551	236	390	367	195	388	60	449	185					
P1S3	2	216	162	147	284	93	290	281	372						
P1S4	3	117	389	191	634	487	354	566	347						
P1S5	3	457	542	426	128	417	341	436	420						
P1S6	3	219	258	542	617	586	294	533	294	68					
P1S7	3	388	621	575	477	433	506	518	357	579	524				
P1S8	4	301	330	539	276	720	777	1187							

Clip ID	Engagement	Fixation Duration
P1S1	3	368
P1S1	3	435
P1S1	3	336
P1S1	3	517
P1S1	3	283
P1S1	3	243
P1S1	3	492
P1S1	3	124
P1S1	3	286
P1S1	3	325
P1S1	3	132
P1S1	3	122
P1S1	3	695
P1S1	3	269
P1S2	3	551
P1S2	3	236
P1S2	3	390
P1S2	3	367
P1S2	3	195
P1S2	3	388
P1S2	3	60
P1S2	3	449
P1S2	3	185
P1S3	2	216
P1S3	2	162
P1S3	2	147
P1S3	2	284
P1S3	2	93
P1S3	2	290
P1S3	2	281
P1S3	2	372
P1S4	3	117
P1S4	3	389
P1S4	3	191
P1S4	3	634
P1S4	3	487
P1S4	3	354
P1S4	3	566
P1S4	3	347
P1S5	3	457
P1S5	3	542
P1S5	3	426
P1S5	3	128
P1S5	3	417
P1S5	3	341
P1S5	3	436
P1S5	3	420
P1S6	3	219
P1S6	3	258
P1S6	3	542
P1S6	3	617
P1S6	3	586
P1S6	3	294
P1S6	3	533
P1S6	3	294
P1S6	3	68
P1S7	3	388
P1S7	3	621
P1S7	3	575
P1S7	3	477
P1S7	3	433
P1S7	3	506
P1S7	3	518
P1S7	3	357
P1S7	3	579
P1S7	3	524
P1S8	4	301
P1S8	4	330
P1S8	4	539
P1S8	4	276
P1S8	4	720
P1S8	4	777
P1S8	4	1187

Figure 3-10. The duration of fixations was computed from the aggregated data. (left) Same as Figure 3-7 (right) to show the duration of each fixation per clip. (right) The pair of data (engagement, the duration of fixations) used in the ANOVA test.

In each pair of data, the level of engagement $l \in \{1, 2, 3, 4\}$ was a rating score converted from human annotation and selection results; fixation duration was calculated as the length of one fixation using its timestamps. The method of calculating the duration of each fixation was discussed above and measured in milliseconds. However, human annotation and selection results were engagement scores for a 10s clip, not for a fixation. The way to give a rating score for a particular fixation was to find the clip that the fixation belongs and take the rating score of adult engagement from that clip. For example, Figure 3-10 (left, the first row of data) showed that 14 fixations were identified within a clip (ID: P1S1) rated as engagement level 3, with a duration of 368ms, 435ms etc. When using for the statistical analysis, 14 pairs of data was captured as shown in Figure 3-10 (right), and the engagement level of the 14 pairs was the rating score for this clip (level 3), like (3,368) and (3,435) etc.

Firstly, a descriptive statistical analysis shows the total fixation duration and the overall average fixation duration per clip according to engagement levels. The total fixation duration

was calculated from all selected recordings and it increased when the level of engagement increased from level 1 to 3, while it decreased during level 3 to 4 (see Table 3-7). Like the total number of fixations, the distribution of the total fixation duration was also related on the unbalanced distribution of clips in terms of engagement. For example, the highly-engaged category (level = 3) has the longest fixation duration because clips labelled as highly-engaged are the most frequently occurring (40.18%). Combined with the number of fixations, the overall mean fixation duration (MFD, = total fixation duration / total number of fixations) was 368.82ms (SD = 232.66) from all selected clips. The mean fixation duration per clip in each engagement level was shown in Figure 3-11 and it increased according to the increased level of engagement. The longest MFD was measured in the fully-engaged category (468.6ms in level = 4), while the shortest one was at the not-engaged category (243.4ms in level = 1).

Level of Engagement	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Clip Duration (%)	460000ms (13.49%)	930000ms (27.27%)	1370000ms (40.18%)	650000ms (19.06%)
Total Fixation Duration (%)	154802.4ms (9.93%)	376022.4ms (24.11%)	732393.6ms (46.97%)	296155.2ms (18.99%)

Table 3-7. The overall distribution of the engagement levels and the total fixation duration, shown in counts and in parenthesis in percentages.

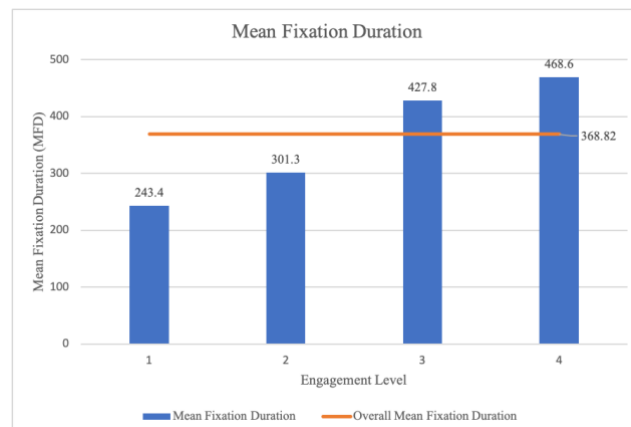


Figure 3-11. The overall distribution of the engagement levels and the average fixation durations. The line shows the overall mean fixation duration of all fixations.

A short summary of the descriptive statistics indicates that the total fixation duration has a same distribution as the total number of fixations, which was also related on the unbalanced distribution of clips in terms of engagement. The average fixation duration per clip increased when the engagement levels increased from level 1 to 4.

A one-way ANOVA was then performed to test for the fixation duration under the effect of the corresponding level of engagement. The test modelled the differences in the mean of the response variable, fixation duration, as a function of the level of engagement. It indicated that there was a statistically significant difference in fixation duration according to the different levels of engagement ($F(3, 4224) = 14.544, p < .001$ under a significant level 0.05). A Hochberg²⁰ *post hoc* test was conducted to examine pairwise comparisons between each engagement level.

The results of the *post hoc* tests show the pairwise differences between the mean value of fixation durations for the independent variable (engagement level), the p-value and its standard error. Table 3-8 shows that the mean differences in fixation duration between all the engagement levels were statistically significant. Therefore, fixation duration had a significant difference among the four engagement levels and is a good indicator for adult engagement classification in the next section.

Dependent Variable: Fixation duration				
Engagement Level l	Engagement Level l'	Mean Difference ($l - l'$)	Std. Error	p-value
1	2	-57.900*	10.624	.000
	3	-184.401*	10.126	.000
	4	-225.249*	12.248	.000
2	1	57.900*	10.624	.000
	3	-126.502*	8.116	.000
	4	-167.351*	10.646	.000
3	1	184.401*	10.126	.000
	2	126.502*	8.116	.000
	4	-40.848*	10.149	.000
4	1	225.249*	12.248	.000
	2	167.351*	10.646	.000
	3	40.848*	10.149	.000

Table 3-8. Results of a Hochberg *post hoc* test to find the actual differences of the four engagement levels on fixation durations. *: The mean difference is significant at the 0.05 level.

In summary, the results of the ANOVAs using two fixation metrics separately, and their *post hoc* tests, show that fixation duration was a better indicator for engagement measurement than the number of fixations per clip because 1) there was a statistically significant difference in fixation duration according to the different engagement levels; 2) there were statistically

significant mean differences in fixation duration with pairwise comparisons between all the engagement levels. While considering the number of fixations per clip, only the fully-engaged category (level 4) had a statistically significant difference between the other three levels. Moreover, the fixation features (e.g., the average and the total fixation duration) were used to further prove that fixation duration contained more information related to different engagement levels than the overall number of fixations per clip (The average fixation duration increased according to the increased level of engagement). Moreover, from the literature, longer fixation duration means that participant's gaze behaviour changes from a slow-paced information extraction to a higher comprehension, as a good indicator of engagement during following of the story video on screen. Thus, fixation duration will be used to classify whether a participant was engaged or not in the next section.

3.4.2 Classification

A 2-class classifier (high vs. low engagement) was built by taking into account the fixation durations to detect whether an adult is engaged in the digital story-stem vignettes. Firstly, a baseline classifier was built that assigned one class which was the most common label. The most common label was the class 'High Engagement' in the dataset. The accuracy of the baseline classifier was 59.24% which means the percentage of clips in the class 'High Engagement' from human observation results is 59.24%.

The accuracy of engagement identification was firstly calculated as it is the most common metric to report classification performance. In the 2-class classification, the accuracy of the classifier was 65.69% (67/102 clips classified successfully). Since the baseline shows that the amount of "High Engagement" class data was greater than the "Low Engagement" class data, the classifier that predicted the most frequent class had a deceptively high accuracy. To prevent this problem, previous studies have used various accuracy to test the performance of the classifier, such as sensitivity/recall (true positive rate) as well as specificity (true negative rate) [48], balanced accuracy [94], F₁ score [80], and Matthews correlation coefficient (MCC) [36, 88]. The metrics used to evaluate the adult engagement classifier were shown in Table 3-9.

Metrics	Description
Precision	A measure of a classifier exactness
Specificity and Sensitivity/Recall	A measure of a classifier completeness
Balanced Accuracy	A measure of overall performance of a classifier without worrying about the imbalance of a dataset
F ₁ score	A weighted average of precision and recall, where an F ₁ score reaches its best value at 1 (perfect precision and recall) and worst at 0.
Matthew Correlation Coefficient (MCC)	A correlation coefficient between the observed and predicted binary classifications; a coefficient of +1 represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and observation.

Table 3-9. Description of accuracy metrics.

A confusion matrix was firstly calculated as shown in Table 3-10 and it shows that 40 clips were correctly classified as ‘High Engagement’ and 27 clips were classified as ‘Low Engagement’. Then accuracy metrics (see Table 3-11) was calculated using the confusion matrix. It shows that the classifier using fixation duration measured adult engagement levels correctly in 67.27% of instances with a balanced accuracy, a good performance for this classification task. Moreover, the classifier using fixation duration has a better performance of exactness than completeness, which means that more clips that correctly labelled were labelled as high engagement.

Prediction \ Annotation	High	Low
	High	40
Low	23	27

Table 3-10. Confusion matrix of the binary classifier of adult engagement (high vs low) using the fixation duration, shown in the number of classified clips.

Method	Acc.	BAcc.	Prec.	Spec.	Sens.	F ₁ score	MCC
Fixations	0.6569	0.6727	0.7843	0.7105	0.6349	0.7018	0.3347

Table 3-11. Accuracy metrics of the binary classification of adult engagement using fixation durations.

3.5 Discussion and Conclusions

This chapter investigates how to use gaze data for measuring adult engagement levels while they are watching digital story-stems. The gaze data were analysed by combining external observation with automated measures to measure adult engagement levels. A system was created to measure the engagement levels of MCAST participants. The work in this chapter

was to develop and test the experimental framework needed for Chapter 4. This work provides foundations for RQ1. Two specific questions are asked:

Q1: What kinds of facial behaviours can be used for designing a coding system for adult engagement levels in story-stem vignettes taken from MCAST?

Human labellers were asked to independently view and rate each 10-second clip based only on the participants' facial appearance. They were asked to annotate each clip with an engagement level (1~4) and note examples of engaged behaviours on a rating form. From the labeller's notes, eye closure was always recorded as a not-engaged behaviour while eyebrow movements, such as eyebrow raise, were recorded as highly-engaged behaviours or even fully-engaged behaviours. The most five frequent facial behaviours that labellers recorded were eyebrow raise, frown, eye closure, mouth/lip movements (e.g. "sad mouth") and laugh (e.g. embarrassed laugh). From this, the facial behaviours related to engagement focuses on movements of eyebrow, eye and lip. These recorded facial behaviours provided a support for designing an annotation scale for the engagement levels, which would be used for labelling children's engagement in the next chapter.

Q2: What features of eye movement data should be analysed for different engagement levels?

Fixation is the main measurement used in this chapter as it is the most commonly used eye-tracking feature. There were three steps for analysing fixations according to different engagement levels. The first step was to identify fixations from the raw gaze data using the velocity-based identification algorithm. Fixation identification was a statistical description of observed eye movement behaviours. There were several default parameters based on previous studies used for improving the accuracy of fixation identification.

The second step was to analyse whether fixation metrics contained information related to distinct levels of adult engagement in digital MCAST story-stems. Firstly, a descriptive statistics of the overall number of fixations and fixation duration were analysed respectively. The total number of fixations and total fixation duration has the same distribution as the distribution of clips in each engagement level, which was unbalanced (an increase from level 1 to 3 but a decrease from level 3 to 4). That is, the total number of fixations and total fixation duration increased during the engagement level 1 to 3 while decreased during level 3 to 4. Meanwhile, when the engagement level increased from level 1 to 4, the average number of fixations decreased (not much difference between level 1 and 2) while the average fixation

duration per clip increased. Then, two one-way ANOVAs using two fixation metrics separately, and their *post hoc* tests, show that there were significant differences in fixation duration across the four engagement levels while the overall number of fixations per clip only in level 4 was different from that in the other engagement levels (1~3). Therefore, the statistical result showed that fixation duration was a better indicator than the overall number of fixations per clip for engagement measurement.

The third step is to evaluate if fixation metrics contain information related to levels of engagement. An SVM classifier using the fixation duration feature classifies engagement correctly in 74.05% of cases with a balanced accuracy, which is a good result of automatic engagement recognition.

Moreover, besides two specific questions, the possible effect of merging clips was also discussed. As recordings of each story-stem cannot be divided as 10-second clips evenly; recordings of each participant were merged together in such a way that the end of one and the start of another end up in one 10s clip. Thus, each participant's recordings can be divided into 18 clips evenly (ID: S1~S18). There were 3 merged clips in one participant's recordings (S3, S6 and S10). The possible effect of merging these clips together was discussed in two aspects: the average rating scores in terms of engagement compared with other clips; and the agreement of rating scores in these merged clips. Firstly, the annotation results showed that relatively low engagement levels were more frequently occurring when rating these merged clips. However, the reason of low engagement rating scores was the nature of story-viewing experience rather than the effect of merging the clips together. The content of MCAST story-stems aims to represent a 'distress' situation with a gradual increase in engagement so that it is acceptable with a relatively low engagement at the beginning of the story-stem. while at the end of story-stems, there was a question that hands over of initiative to the participant that triggers the next phase. That trigger created a decreasing engagement to ensure that the participant could have a smooth transition of initiative in commencing narrative. Secondly, the agreement of rating scores for these merged clips was acceptable. For example, the ratings of merged clips for participant P1 has a moderate agreement, with 66.7% of clips. Thus, in this study it is important to understand the nuances of the engagement development to design a coding system that could be used for child engagement measurement. The merging clips could provide more data, specifically on 'disengaged' examples, to have a better understanding of low engagement levels.

Therefore, this chapter is a *preliminary test* for Chapter 4 and there are two main contributions: 1) to develop an annotation scale based on adults' engagement behaviours; 2) to analyse gaze behaviours using automated engagement measurements. The annotation scale will be used for coding children's engagement levels in Chapter 4 to measure their engagement levels. Fixation duration as a good indicator of adult engagement measurements will be used to compare to children's gaze behaviours to see if they are similar. Additionally, the effect of merging together the clips in such a way that the end of one and the start of another end up in one 10s clip was also discussed. The way of merging clips was acceptable based on a moderate agreement of rating scores so that more data could be used to measure adult engagement levels. The next chapter will focus on child engagement measurements in the digital MCAST story-stems.

Chapter 4 Child Engagement

Measurements from Facial Data

4.1 Introduction

Engagement has been recognised as a key factor in understanding children’s psychology and behaviour that has made significant contributions to understanding attachment with caregivers [15, 32, 67] in the context of story-stems. In the story-stem approach, an interviewer gives the beginning of a story then asked the child to complete it, often acting out the scene using dolls. The instance of the story-stems approach used in this thesis is the widely-used Manchester Child Attachment Story Task (MCAST), as discussed in Section 2.3. Engagement is important in the story-stem, where it is in initial phase of the test – a child is given the beginning of a story by an assessor using two dolls. According this, child engagement in this thesis is defined as *a focusing of children’s mood state around the particular distress represented in the MCAST story-stem, which means that children focus on attending to the play and materials, are not distracted by other things, and feel empathy with the dolls and characters in the story.*

The problem of evaluating child engagement in the MCAST test has motivated great interest in methods to measure it. In the traditional MCAST test, engagement is measured by a trained assessor’s observation of facial expressions, using the MCAST protocol. If a child is not emotionally engaged by the predicament shown in each story-stem, then the test will not be successful and the data collected will not allow for an MCAST assessment related to child Attachment status. Unfortunately, as Section 2.3 described, conducting MCAST assessments is expensive and time-consuming. In order to reduce the time and cost required for engagement assessment, a system called SAM has been developed to automated Attachment assessment. However, the SAM system itself does not detect the child’s engagement. Therefore, this chapter focuses on measuring children’s engagement levels in digital SAM story-stems, which investigates the answer to RQ1: *Can children’s spontaneous facial expressions be used to automatically measure engagement levels in digital story-stems?*

Chapter 3 has shown that adult engagement can be measured using gaze behaviours based on human annotation. Besides using the external observation and the eye-tracking measures taken from Chapter 3, this study also uses the self-report as the third measurement method as well as facial expression recognition as another automatic measure. The self-report measure could provide a participant's perspective of a system based on his/her mental state to help researchers understand the participant's engagement. This study designs a Smiley-o-meter questionnaire [73] that focuses on children's mental state, such as attention and emotion, to interpret their engagement levels. However, researchers have argued that the questionnaire may not be suitable for all users because the accuracy of users' answers relies on their interpretation of the questions and the person's feelings at the time they filled out the questionnaire. Hanna and colleagues suggested that children's observed facial expressions could be a better engagement indicator than their answers to questionnaires [34]. Children's answers of the questionnaire will be used to compare the results of external observation. Besides eye-tracking, Section 2.5.5 shows that facial expression recognition is another automatic method for measuring children's engagement levels because child engagement is measured by human observations of facial expressions in the original MCAST assessment [32]. However, there has been few studies of automatic child engagement measurements through analysing cues from the face and gestures. Children's facial expressions in this study are coded using the Facial Action Coding System (FACS) [25] as introduced in Section 2.5.5, which displays the intensity of over 40 distinct facial muscles around the eyes, eyebrows, and upper cheeks in terms of individual facial Action Units (AUs). This chapter aims to detect if the FACS could be used to measure children's engagement levels in the context of story viewing based on human annotation. The annotation scale is taken from Chapter 3.

Therefore, this chapter aims to answer the RQ1 and investigates the performances among these three measurement methods. In this experiment, facial data from 20 children are recorded by an RGB video camera for extracting facial expressions, along with a Tobii eye tracker for extracting gaze behaviours, while watching the digital story-stems. The engagement levels of each child were then manually coded. From this, a statistical analysis of facial data was presented across different levels of engagement from which several face features were extracted and used to classify the engagement level of children.

4.2 Methods

4.2.1 Participants

Twenty children (7-10 years old, 10 males and 10 females) were recruited from several Glasgow (UK) schools based on their school and parental agreement. After a school agreed to participate, classroom teachers sent opt-out consent forms to each child's family. The forms are shown in Appendix D. This informed families about the research project, explaining the research into a better understanding of child engagement in digital story-stems, introducing the types of data to be collected and the tools to be used for data collection. In the event that a child's family opted out of participating in the research, the child was not selected to participate.

The MCAST usage is for middle-aged children (5-10 years old). This study needs children to fill in a smiley-o-meter questionnaire as shown in Appendix E about their story experience after they watched the digital story-stem. A pilot testing revealed that young children (5-6 years old) had difficulty understanding the questionnaire. Therefore, after discussing with MCAST experts, a slightly older group (7-10 years old) were recruited for the study. This group would still be suitable for MCAST as it is used on children of this age in practice.

4.2.2 Procedure

The test took approximately 20 minutes for each child. To start, an introduction to the procedure and play materials including two dolls and a doll's house with furniture was given. The child's eye movements were calibrated using the Tobii's calibration procedure¹⁰. The four 'distress' SAM story-stem vignettes (see Section 2.3) were then presented. For each of 'distress' scenario, there is an induction phase where the child is given the beginning of a story that is represented on the computer by a short animation narrated by a human storyteller's voice. Figure 4-1 and Figure 4-2 are two screenshots taken from two MCAST stories used in the SAM system. The second phase of the vignette, the child plays out a story to completion with the materials available. After the story-completion task, the administrator asks the child to fill in a smiley-o-meter questionnaire.

During the watching session, the child's facial expressions were collected by a Logitech C920¹¹ web camera and eye gaze data were collected using a Tobii EyeX¹² eye-tracker as in Chapter 3. The camera was placed the same as Chapter 3. Methods for establishing temporal

correspondence between the two recording systems (i.e. the eye-tracking signals and webcam video) were introduced in Section 3.2.2.



Figure 4-1. A screenshot of the 'nightmare' story-stem.



Figure 4-2. A screenshot of the 'illness' story-stem.

4.2.3 Data Annotation

The recorded videos of children watching the story-stem vignettes were then labelled in terms of child engagement. The five labellers (L1~L5) who were recruited in Chapter 3 were still asked to label the recorded videos. Section 2.5.2 provided two timescales for video annotation for engagement levels for the two automated measures (i.e., eye-tracking and facial expression recognition) respectively. For eye-tracking, the annotation procedure was the same as Chapter 3 (see Section 3.2.3). There were 360 clips in total and each clip was annotated by two labellers. Each labeller was allocated to rate 144 clips and they were asked to finish the rating in 5 days.

Since facial expression recognition was based on the Facial Action Coding System (FACS), which means that information of facial action units could be more easily captured from static images than short video clips, recordings were split into static frames for facial expression recognition then given a single number to rate engagement for each frame. All recordings were split into static images. Each one second clip was split into 30 frames giving 108000 frames in total (60mins of recordings). Five labellers (L1~L5, same as labellers recruited in Chapter 3) were asked to rate the frames like rating the clips and all frames were randomly allocated to the labellers.

Level	Name	Characteristic
1	Not engaged	e.g. looking away from screen and focusing on something other than the story; eyes completely closed over 3 seconds
2	Rarely engaged	e.g. clearly not “into” the story; paying attention to something else (e.g. camera and desktop eye-tracker), but sometimes focusing on the story
3	Highly engaged	e.g. good enough to proceed to the task such as fixed eyes on the screen; participant requires no admonition to “stay on task”
4	Fully engaged	e.g. good quality engagement such as keep gaze on the screen; participant could be “commended” for his/her level of engagement in task
X		The frame was very unclear or contains no person at all.

Table 4-1. The engagement level annotation categories used by the raters.

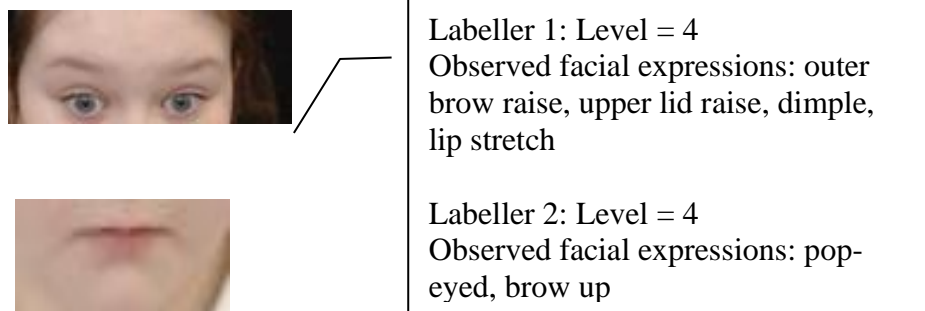


Figure 4-3. An example of human annotation using a frame.

Given the approach of rating a single engagement level $l \in \{1, 2, 3, 4\}$ for each image, the annotation scale was taken from Chapter 3 as shown in Table 4-1 to distinguish four different levels of child engagement, ranging from the not engaged to the fully engaged category. If one clip was unclear (e.g. no eyes, eye/face occlusion) or contains no person at all, they were asked to annotate this clip with an X. For example, Figure 4-3 shows a screenshot²¹ of one

²¹ In order to protect the personal information, only part of the screenshot of the participant's recording was shown here.

participant during the story watching and its annotation result and notes of the child's facial expressions from two labellers.

4.2.4 The Inter-Rater Reliability

Pilot Testing using the Fleiss Kappa

To detect if this scale could also be suitable for annotating child engagement levels, a pilot test was undertaken using chosen 40 clips and 400 frames. These clips and frames were taken from the SAM study with permission. Also, the SAM assessors were asked to rate the selected clips and frames in terms of children's engagement levels $l \in \{1, 2, 3, 4\}$. In order to test the agreement of data annotation in each level of child engagement, the distribution of clips/ frames was balanced. There were 10 clips as well as 100 frames in each engagement level using the SAM result. The agreement across human labellers was calculated. Since the number of human labellers was more than two, the inter-rater reliability was performed using a Fleiss' kappa²² rather than a Cohen's kappa. Meanwhile, the scores of each clip was then calculated by rounding the average score for that image to the nearest integer (e.g., 2.4 rounds to 2; 2.5 rounds to 3). The average score of each clip was used to compare to the engagement score taken from the SAM test.

- 1) Annotation of clips: Fleiss' kappa was 0.450 (s.e. = 0.0290) under a 95% confidence interval across the 5 labellers, which can be recognised as a moderate agreement for annotating children's engagement levels. The agreement of each level of child engagement was also calculated, as shown in Table 4-2.

For comparing to the SAM result, the rating result of five labellers showed that only one frame has a different score with the SAM rating scores (The average score from the five labellers was 2 while the SAM rating score in that clip was 1).

Level of Engagement	Total	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Fleiss Kappa	0.449045	0.469534	0.398119	0.381498	0.553994

Table 4-2. Results of the agreement of annotating the clips in terms of child engagement using Fleiss kappa.

²² Fleiss' kappa (named after Joseph L. Fleiss) is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items.

- 2) Annotation of frames: Fleiss' kappa was 0.437 (s.e. = 0.0091) under a 95% confidence interval across the 5 labellers, which can be recognised as a moderate agreement for annotating children's engagement levels. The agreement of each level of child engagement was also calculated, as shown in Table 4-3.

For comparing to the SAM result, the rating results from five labellers showed that 89% of frames had the same rating score as the SAM result. The distribution of frames with different ratings between the average scores from the five labeller and the SAM ratings was 10 (10%), 11 (11%), 16 (16%) and 7 (7%) from level 1 to 4 respectively.

Level of Engagement	Total	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Fleiss Kappa	0.436368	0.547051	0.343236	0.290949	0.549851

Table 4-3. Results of the agreement of annotating the frames in terms of child engagement using Fleiss kappa.

In this study using the Cohen's kappa

The result of the pilot test showed that there was a moderate agreement of recognising children's engagement levels both using clips and frames. This indicated that the five labellers had an agreement of child engagement annotation and their accuracy of annotation was also acceptable. The next step was to calculate the agreement of human annotation using clips and frames recorded in this study respectively. Like Chapter 3, the inter-rater reliability, the degree of agreement between the two labellers, was performed using a weighted Cohen's κ . The weighted kappa is calculated using a pre-defined table of weights which measures the degree of disagreement between the two independent labellers, the higher the disagreement the higher the weight.

- 1) Annotation of clips

After the data annotation, there were 12 clips (3.33%) that labelled as X with agreement due to bad quality and eye/face occlusion and bad quality. The distribution of annotation scores was shown in Table 4-4 (left).

Cohen's κ was 0.793 (s.e. = 0.026) in this dataset, which means that labellers have a substantial agreement (0.61~0.80) in rating child engagement levels. The percentage of agreement, which the proportion of clips with the two same scores, was $293/348 = 84.19\%$. There were 246 clips (70.69%) that were labelled as 'engaged', including level 3 and 4. It

indicated that most children were engaged during the watching session. Cohen's κ in the 'engaged' dataset was 0.802 (s.e. = 0.042), which means that labellers have an almost perfect agreement (0.81~1.80) in labelling the engaged data. The proportion of clips which the two scores were the same was 91.87% over all engaged clips.

		Labeller 1 Engagement Score				
		1	2	3	4	Total
Labeller 2 Engagement Score	1	14	8	2	0	24
	2	5	53	8	0	66
	3	1	9	165	11	186
	4	0	2	9	61	72
	Total	20	72	184	72	348
Agreement		14	53	165	61	293

0	1	2	3
1	0	1	2
2	1	0	1
3	2	1	0

The table of weights.

Table 4-4. (left) Overall Child engagement ratings (12 clips labelled as X due to quality). (right) The table of weights. The agreement across child engagement levels 1, 2, 3, and 4 agreement 84.19%: Cohen's kappa = 0.793. Engaged data (level 3 and 4) agreement 91.87%: Cohen's kappa = 0.802.

2) Annotation of frames

After the data annotation, there were 4018 frames (3.72%) that labelled as X with agreement due to eye/face occlusion and bad quality. The distribution of annotation scores was shown in Table 4-5 (left).

Cohen's κ was 0.710 (s.e. = 0.002) in this dataset, which means that labellers have a substantial agreement (0.61~0.80) in rating the engagement levels. The percentage of agreement, which the proportion of clips with the two same scores, was $76560/103972 = 73.64\%$. There were 75999 frames (73.09%) that were labelled as 'engaged', including level 3 and 4. The result indicated that most children were engaged during the watching session. Cohen's κ was 0.561 (s.e. = 0.003) in this 'engaged' dataset, which means that labellers have a moderate agreement (0.41~0.60) in labelling the engaged data. The proportion of clips which the two scores were the same was 78.52% over all engaged clips.

		Labeller 1 Engagement Score				
		1	2	3	4	Total
Labeller 2 Engagement Score 2	1	5000	3408	664	2	9074
	2	3529	11882	1087	158	16656
	3	533	1491	35501	8028	45553
	4	8	211	8293	24177	32689
	Total	9070	16992	45545	32365	103972
Agreement		5000	11882	35501	24177	76560

0	1	2	3
1	0	1	2
2	1	0	1
3	2	1	0

The table of weights.

Table 4-5. (left) Overall child engagement ratings (4108 frames labelled as X due to quality). (right) The table of weights. The agreement across child engagement level 1, 2, 3, and 4 agreement 73.64%: Cohen's kappa = 0.710. Engaged data (level 3 and 4) agreement 78.52%: Cohen's kappa = 0.561.

Moreover, comparing to the agreement of annotating the clips, the reliability of agreement for labelling the frames has a lower kappa value, specifically on labelling the 'engaged' data (0.802 vs. 0.561). It indicated that human labellers had not a good performance of distinguishing 'highly-engaged' and 'fully-engaged' in annotating a single static image than as annotating a 10s clip. Thus, the next step was to determine how to select the data with a better reliability based on the two different annotation results.

4.2.5 Data Selection

Once the data had been annotated, training and testing data were selected for classification. There were two selection procedures corresponding to two timescales used for annotating children's engagement levels. Firstly, the selection procedure using gaze data for analysis and classification was the same as the procedure from Chapter 3 (see Section 3.2.4). While selecting training and testing data related to facial AUs for classification, due to the large number of frames as well as the moderate agreement of labelling the 'engaged' data, only frames with two same scores were retained. Otherwise, the frame was discarded.

After data selection, the distribution of engagement was shown in Table 4-6 and Table 4-7 according to the two selection procedures. The next step was to connect the annotation of

children's engagement to the engaged behaviours using two automated measurements respectively to classify the engagement levels.

4.3 Recognition of Child Engagement

4.3.1 Recognition using Eye-tracking

All recordings (60mins of recordings) were split into 10-second video clips giving 360 clips in total. After the procedure of data annotation and selection, the total number of annotated clips were 348 (12 clips were labelled as X and discarded.) The number (percentage) of clips in each level of child engagement is 23 (6.61%), 61 (17.53%), 189 (54.31%), and 75 (21.55%) from level 1 to level 4 respectively (see Table 4-6).

Level of Engagement	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Count of clips (%)	23 (6.61%)	61 (17.53%)	189 (54.31%)	75 (21.55%)

Table 4-6. The distribution of clips in terms of child engagement levels, shown in the count of clips and in parenthesis in percent.

The procedure of recognition of child engagement using eye-tracking measures was the same as the procedure from Chapter 3 (see Section 3.3), including grouping the gaze data into fixations, computing fixation metrics and conducting a classification task.

4.3.2 Recognition using Facial AUs

All recordings were split into static images. Each one second clip was split into 30 frames giving 108000 frames in total (60mins of recordings). After the procedure of data annotation and selection, the total number of annotated frames were 76560 (4018 frames were labelled as X and 27422 frames were labelled with two different rating scores. Both were discarded.) The number (percentage) of frames in each level of child engagement is 5000 (6.53%), 11882 (15.52%), 35501 (46.37%), and 24177 (31.58%) from level 1 to level 4 respectively.

Level of Engagement	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Count of frames (%)	5000 (6.53%)	11882 (15.52%)	35501 (46.37%)	24177 (31.58%)

Table 4-7. The distribution of frames in terms of child engagement levels, shown in the count of clips and in parenthesis in percent.

Facial AUs Recognition

Facial features were extracted from the 76560 frames that were selected using the data selection procedure. OpenFace [9] was employed for extracting facial features in this study, which is a fully open source real-time facial behaviour analysis system using Support Vector Machines (SVM) for AU occurrence detection and Support Vector Regression (SVR) for AU intensity detection [8]. For facial AU recognition, it is able to recognise a subset of AUs, specifically: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, and 45 (see Figure 2-5). The intensity and presence of each facial Action Unit would be used to measure children's engagement levels during the MCAST test.

Classification

Children's facial data were annotated by human labellers with four classes. As the classifier aims to detect whether a child is engaged or disengaged in the digital stories, a 4-class classification task was turned into a classification problem with two classes: low engagement levels and high engagement levels. Class A was the low engagement levels including the not engaged and the rarely engaged categories while the highly engaged and fully engaged categories were grouped into Class B for high engagement levels. To perform the classification task, the LIBSVM²³ library [19] was used as an efficient implementation of the standard soft-margin Support Vector Machine (SVM) [20]. In the first step, data were normalised by linear transformation into an [0,1] interval. Second, the SVM classifier was created to detect whether a child is engaged in the digital story-stem vignettes from the extracted facial action units. In designing classifiers, frames with valid facial AUs was divided into two sets: one with 70 percent of the source data, for training the model, and one with 30 percent of the source data, for testing the model.

4.4 The Self-report Measure

A questionnaire instrument is the most commonly-used technique for the self-report measure in engagement from prior research [23]. There were seven single choice questions (Q1- Q7) using a smiley-face based 5-point Likert scale [73]. Child participants were asked to give a rating of to each question by choosing a smiley face. Q8 and Q9 were two open-ended questions at the end, which support children to describe their story experience and attitudes. All questions and its related aspects were shown in Table 4-8. As discussed in Section 2.5.1,

²³ www.csie.ntu.edu.tw/~cjlin/libsvm/

this questionnaire aims to measure child engagement that consists of four aspects: distraction/attentional focus, empathy, story understanding and general attitude towards this task/interest. Distraction, also called attentional focus, is a concentration of mental activity; it means that concentrating on the story-stem only and ignoring all other things. Empathy means children's feeling with the character's emotions, like the child doll. Story understanding requires a child located him or herself within the mental models of the story-stem. General attitude towards this task/Interest focuses on children's feelings of being interested and having fun during the story watching. Full description of the aspects can be seen in Table 2-4 and the aspect of aesthetics will be investigated in Chapter 5.

Questions	Aspects
Q1. I was absorbed in this story.	Distraction/attentional focus
Q2. I was involved in this story that I'm happy to tell people what happens next.	Distraction/attentional focus
Q3. I found this story confusing to understand.	Story understanding
Q4. When I was watching this story, I found myself thinking about other things.	Distraction/attentional focus
Q5. I was stressed while watching this story.	Empathy
Q6. I felt I knew what the child doll were going through emotionally in this story.	Empathy
Q7. I felt interested in this story task (including the story and this questionnaire).	General attitudes towards this task/ Interest
Q8. How's the mummy doll feeling now? And what's the mummy doll thinking now?	Story understanding + Empathy
Q9. How's the child doll feeling now? And what's the child doll thinking now?	Story understanding + Empathy

Table 4-8. The items of the questionnaire and its related aspects.

4.5 Results of Eye-tracking Measures

There are two ways taken from Chapter 3 to measure children's engagement levels using the selected clips when children were watching the digital MCAST story-stems. Firstly, there was a statistical analysis using fixation metrics including the overall number of fixations and fixation duration with regard of children's different engagement levels based on results from human annotation. A classification using children's fixation metrics is then performed where each clip is marked as being either low or high engagement.

4.5.1 Primary Fixation Metrics

This section presents a statistical analysis of the two fixation metrics with regard to the different levels of child engagement. Twenty child participants (recorded as P1~P20) performed a total of 80 story-stem trials (4 MCAST story-stems for each child). Like the data collection procedure in Chapter 3, the video recordings for children were split into 10-second video clips (18 clips per child, recorded as S1~S18) and the raw data were collected by an eye tracker and each gaze data point was identified by a timestamp and (x, y) coordinates (see Figure 4-4 (top), gaze data was taken from a 10-second clip with ID: P1S1). The coordinate system used here and the timestamp were introduced in Section 3.4.1.

Aggregating data was performed to group the raw gaze data into fixations and calculate the duration of fixations to record them. The methods of grouping the gaze data into fixations as well as calculating the duration of each fixation were discussed in Chapter 3 and the duration of identified fixations was measured in milliseconds. Figure 4-4 (bottom) shows the duration of each identified fixations in each clip. For example, the row with clip ID P1S1 included 11 fixations and the duration of the first fixation was 141ms.

Clip ID	Gaze Data		Timestamp	
P1S1	915.2	727.6	timestamp	16057435ms
P1S1	911.4	722.6	timestamp	16057465ms
P1S1	939.9	715.7	timestamp	16057480ms
P1S1	939.3	715	timestamp	16057525ms
P1S1	938.1	711.2	timestamp	16057540ms
P1S1	934.7	710.4	timestamp	16057572ms
P1S1	932	711	timestamp	16057585ms
P1S1	929.9	711	timestamp	16057600ms
P1S1	928	707.8	timestamp	16057616ms
P1S1	933.6	707.8	timestamp	16057646ms
P1S1	938.1	708.8	timestamp	16057660ms
P1S1	941.8	710.1	timestamp	16057675ms
P1S1	943.4	712.3	timestamp	16057690ms
P1S1	942	713.7	timestamp	16057735ms

Clip ID	Engagement	Fixations (recorded by duration(ms))																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
P1S1	3	141	320	402	312	85	154	171	189	135	255	226								
P1S2	4	329	370	503	332	427	478	487	572	234	422	245	317	323						
P1S3	3	239	256	230	551	231	182	250	139	219										
P1S4	3	244	371	477	456	344	207	108	118	127	379	194	416	432	119					
P1S5	3	156	179	697	232	550	330													
P1S6	3	291	349	458	85	196	244	290	301	319	565	90	195	659	136	106	173	152	137	
P1S7	4	552	572	389	336	730	687	670	375	313	739	608	135							
P1S8	4	589	571	784	480	460	389	414	383	784	363	430	271							
P1S9	4	215	437	517	331	446	531	512	524	667	463									
P1S10	2	156	80	73	151	209	164	173												
P1S11	2	465	198	318	324	276	225	241	158	340	146	188	252	386	404	428	263	259	255	331

Figure 4-4. The screenshots of collected and aggregated data. (top) Raw gaze data was collected by the eye tracker. (bottom) The raw gaze data were grouped into fixations and combined with child engagement ratings from human annotation.

For the column titled ‘engagement’ in Figure 4-4 (bottom), the level of child engagement was taken from the annotation results of video recordings for all children. The number (percentage) of the selected 348 video clips (discarded for 19 clips due to eye/face occlusion) in each level of child engagement was shown in Table 4-6. Therefore, the aggregated data

included three categories: clip ID, child engagement level and fixations (recorded by durations), shown in Figure 4-4 (bottom).

The overall number of fixations

In this section, analysing the number of fixations determined whether this feature contained information related to child engagement levels during watching the story-stem videos. The variables used here were computed as a set of the-number-of-fixations/the-level-of-child-engagement pairs in each 10-second segment, which were recorded in the forms of numbers (see Figure 4-5 (right)). In each pair of data, the level of child engagement $l \in \{1, 2, 3, 4\}$ was a rating score taken from human annotation and selection results; fixations were recorded by their durations (see Figure 4-5 (left), same as Figure 4-4 (bottom)), and the number of fixations was computed per clip. For example, Figure 4-5 (left) shows that the first row included 11 elements in the ‘fixation’ category. This row was converted into the first row (the same clip ID: P1S1) in the right graph to be analysed.

Clip ID	Engagement Fixations (recorded by duration/ms)																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
P1S1	3	141	320	402	312	85	154	171	189	135	255	226								
P1S2	4	329	370	503	332	427	478	487	572	234	422	245	317	323						
P1S3	3	239	256	230	551	231	182	250	139	219										
P1S4	3	244	371	477	456	344	207	108	118	227	379	194	416	432	119					
P1S5	3	156	179	697	232	550	330													
P1S6	3	291	349	458	85	196	244	290	301	319	565	90	195	639	136	106	173	152	137	
P1S7	4	552	572	389	536	730	687	670	375	313	739	608	135							
P1S8	4	589	571	784	480	460	389	434	383	784	363	430	271							
P1S9	4	215	437	517	331	446	551	512	524	667	463									
P1S10	2	136	80	73	151	209	164	173												
P1S11	2	465	198	318	324	276	225	241	158	340	146	188	252	386	404	428	263	259	255	331

Clip ID	Engagement	Number of Fixations
P1S1	3	11
P1S2	4	13
P1S3	3	9
P1S4	3	14
P1S5	3	6
P1S6	3	18
P1S7	4	12
P1S8	4	12
P1S9	4	10
P1S10	2	7
P1S11	2	19

Figure 4-5. The number of fixations per clip was computed from the aggregated data. (left) Same as Figure 4-4 (right) to show the duration of each fixation in one 10s clip. (right) The pair of data (child engagement, the number of fixations) was used in the ANOVA test.

Firstly, a descriptive statistical analysis shows the total number of fixations and the average number of fixations per clip according to four child engagement levels. Overall, the 348 clips contained 4853 fixations and children would averagely have about 14 fixations during a 10-second segment. The total number of fixations increased according to the increased level of child engagement from level 1 to 3, while it decreased from level 3 to 4 (see Table 4-9), corresponding to the number of clips in each engagement level. For example, the highly-engaged category (level = 3) had the highest number of fixations as clips labelled as highly-engaged were the most frequently occurring (54.31%). The average number of fixations per clip (= the total number of fixations / clip counts) was then computed for child engagement level 1 to 4 (see Figure 4-6). The average number of fixations per clip decreased when child engagement increased from level 1 to 4 but there was not much difference between level 3

and 4. Thus, the descriptive analysis indicates that the total number of fixations increased from the engagement level 1 to 3 and decreased from level 3 to 4, which was related on the unbalanced distribution of clips in terms of child engagement. Meanwhile, the average number of fixations per clip increased when the engagement levels increased from level 1 to 4, but there was not much difference in the average number of fixations between the high-engaged and the fully-engaged category.

Level of Child Engagement	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Clip Count (%)	23 (6.61%)	61 (17.53%)	189 (54.31%)	75 (21.55%)
Total number of Fixations (%)	407 (8.39%)	971 (20.00%)	2489 (51.29%)	986 (20.32%)

Table 4-9. The overall distribution of child engagement and the total number of fixations, shown in counts and in parenthesis in percentages.

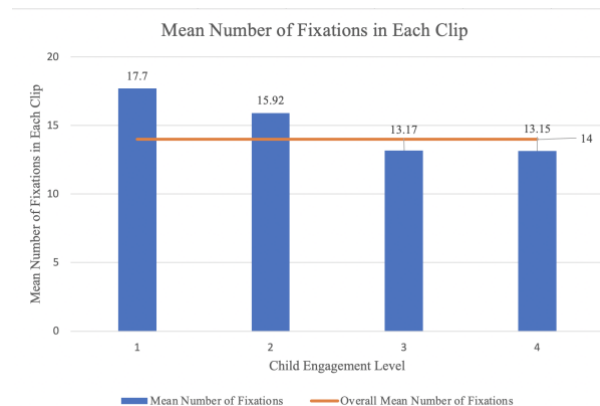


Figure 4-6. The overall distribution of child engagement levels and the average number of fixations per clip. The line shows the overall mean number of fixations of all fixations.

A one-way ANOVA was then performed to test for an effect of child engagement levels on the overall number of fixations per clip. The test modelled the differences in the mean of the response variable, the number of fixations per clip, as a function of child engagement level. It indicates that there was a statistically significant difference in the number of fixations per clip according to the four levels of children's engagement ($F(3, 344) = 4.989$, $p = .002$ under a significant level 0.05). To investigate where the actual differences were in the ANOVA test, a Hochberg²⁰ *post hoc* test was conducted as pairwise comparisons among groups for the independent variable (child engagement level). Table 4-10 shows the differences in the mean number of fixations per clip between child engagement levels, the p-value and its standard error for multiple pairwise comparisons.

The result of the *post hoc* test shows significant pairwise mean differences in the number of fixations per clip across child engagement levels. There was an average difference of 4.526

($p = .020$) between the engagement level 1 and 3; an average difference of 4.459 ($p = .037$) between level 1 and 4; and an average difference of 2.749 ($p = .044$) between level 2 and 3. However, the differences in the number of fixations between other engagement level groups, were not statistically significant. Therefore, the overall number of fixations in one clip is not a good indicator for child engagement classification in the next section as the mean differences in the number of fixations per clip across child engagement levels, were not always statistically significant.

Dependent Variable: The number of fixations in each clip				
Engagement Level l	Engagement Level l'	Mean Difference ($l - l'$)	Std. Error	p-value
1	2	1.778	1.700	.878
	3	4.526*	1.534	.020
	4	4.459*	1.656	.037
2	1	-1.778	1.700	.878
	3	2.749*	1.023	.044
	4	2.771	1.198	.121
3	1	-4.526*	1.534	.020
	2	-2.749*	1.023	.044
	4	.023	.948	1.000
4	1	-4.459*	1.656	.037
	2	-2.771	1.198	.121
	3	-.023	.948	1.000

Table 4-10. Results of a Hochberg *post hoc* test to find the actual differences of the four child engagement levels on the overall number of fixations per clip. *: The mean difference is significant at the 0.05 level.

Fixation Duration

Analysing fixation durations determined whether this feature contained information related to child engagement levels during watching the story-stem videos. The variables was a set of fixation-duration/the-level-of-child-engagement pairs and each pair of data was recorded in the forms of numbers (see Figure 4-7 (right)). All 348 clips contained 4853 fixations so that there were 4853 pairs of data. In each pair of data, the level of child engagement $l \in \{1, 2, 3, 4\}$ was a rating score converted from human annotation and selection results; fixation duration was calculated using its timestamp. Similar to Chapter 3, rating scores in terms of child engagement was annotated for a 10s clip, not for a fixation. Child engagement level to a particular fixation was taken from the engagement level of the 10s clip where the particular fixation belongs. For example, the first row of Figure 4-7 (left) shows 11 fixations were identified within a clip rated as level 3 of child engagement, with a set of durations

141ms and 320ms etc. In this clip, 11 pairs of data was captured, such as (3,141) and (3,320) (see Figure 4-7 (right)).

Clip ID	Engagement Fixations (recorded by duration(ms))																			Clip ID	Engagement	Fixation Duration	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19				
P151	3	141	320	402	312	85	254	171	389	195	255	226								P151	3	141	
P152	4	329	370	503	332	427	478	487	571	134	422	245	317	323						P151	3	320	
P153	3	239	156	230	551	231	182	250	139	219										P151	3	402	
P154	3	244	371	477	456	344	207	108	118	127	379	194	436	432	119					P151	3	312	
P155	3	156	179	697	232	550	330													P151	3	85	
P156	3	291	349	458	85	196	244	290	301	319	565	90	195	659	136	106	173	152	137	P151	3	154	
P157	4	552	572	389	338	730	487	670	375	313	739	608	135							P151	3	171	
P158	4	589	571	784	480	460	389	424	383	784	383	430	272							P151	3	189	
P159	4	215	437	517	331	446	531	512	534	667	463									P151	3	135	
P15D	2	156	80	73	151	209	164	173												P151	3	255	
P1511	2	465	198	318	324	276	225	241	158	340	146	188	252	386	404	428	263	259	255	331	P151	3	154
																					P151	3	171
																					P151	3	189
																					P151	3	135
																					P151	3	255
																					P151	3	226
																					P152	4	329
																					P152	4	370
																					P152	4	503
																					P152	4	332
																					P152	4	427
																					P152	4	478
																					P152	4	487

Figure 4-7. The duration of fixation was computed from the aggregated data. (left) Same as Figure 4-4 (right) to show the duration of each fixation per clip; (right) The pair of data (child engagement, the duration of each fixation) was used in the ANOVA test.

Firstly, a descriptive statistical analysis shows the total fixation duration and the overall average fixation duration per clip according to the four child engagement levels. The total fixation duration in each level of child engagement was shown in Table 4-11 and it increased when the level of child engagement increased from level 1 to 3, while it decreased from child engagement recorded by level 3 to 4. It indicates that the distribution of the total fixation duration was related on the unbalanced distribution of clips in terms of child engagement. For example, the not-engaged category (level = 1) has the shortest total fixation duration as clips labelled as not-engaged are the least frequently occurring (6.61%). Combined with the number of fixations, the overall mean fixation duration (MFD, = total fixation duration / total number of fixations) across all clips was 386.85ms (SD = 206.20). The average fixation duration per clip in each level of child engagement was shown in Figure 4-8 and it increased according to the increased level of child engagement. The longest MFD was measured in the fully-engaged category (457.9ms in level = 4), while the shortest average duration was at the not-engaged category (259.1ms in level = 1).

Level of Child Engagement	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Clip Duration (%)	230000ms (6.61%)	610000ms (17.53%)	1890000ms (54.31%)	750000ms (21.55%)
Total Fixation Durations (%)	105453.7ms (5.62%)	309166.4ms (16.47%)	1011280.7ms (53.86%)	451489.4ms (24.05%)

Table 4-11. The overall distribution of child engagement and the total fixation duration, shown counts and in parenthesis in per cents.

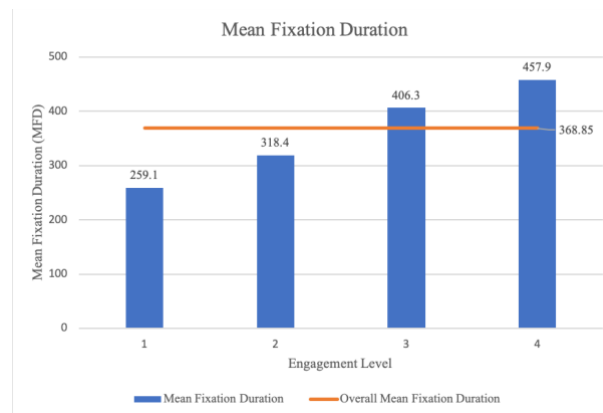


Figure 4-8. The overall distribution of child engagement levels and mean fixation durations. The line shows the overall mean fixation duration in all clips.

The descriptive statistics was similar compared to it in Chapter 3. It indicates that the total fixation duration has a same distribution as the total number of fixations, which was also related on the unbalanced distribution of clips in terms of child engagement. The average fixation duration per clip increased when child engagement levels increased from 1 to 4.

Dependent Variable: Fixation duration				
Engagement Level l	Engagement Level l'	Mean Difference ($l - l'$)	Std. Error	p-value
1	2	-59.249*	11.644	.000
	3	-147.782*	10.561	.000
	4	-198.793*	11.637	.000
2	1	59.249*	10.664	.000
	3	-87.323*	7.473	.000
	4	-139.944*	8.930	.000
3	1	147.782*	10.561	.000
	2	87.323*	7.473	.000
	4	-51.611*	7.433	.000
4	1	198.793*	11.637	.000
	2	139.944*	8.930	.000
	3	51.611*	7.433	.000

Table 4-12. Results of a Hochberg *post hoc* test to find the actual differences between the four child engagement levels on fixation durations. *: The mean difference is significant at the 0.05 level.

A one-way ANOVA was then performed to test for the fixation duration under the effect of the corresponding level of child engagement. The test modelled the differences in the mean of the response variable, fixation duration, as a function of the level of child engagement. It indicated that there was a statistically significant difference in fixation duration according to the different levels of child engagement ($F(3, 4849) = 11.468, p < .001$ under a significant

level 0.05). A Hochberg²⁰ *post hoc* test was conducted to examine pairwise comparisons between child engagement levels. The result of the *post hoc* test shows the pairwise differences in the mean values of fixation duration among groups for the independent variable (child engagement level), the p-value and its standard error. Table 4-12 shows the mean differences in fixation duration between all the engagement levels were statistically significant. Therefore, fixation duration had a significant difference according to the four child engagement levels and is a good indicator for child engagement classification in the next section.

In summary, the results of the ANOVAs using two fixation metrics separately, and their *post hoc* tests, show that fixation duration was a better indicator than the number of fixations per clip for child engagement measurement because 1) there was a statistically significant difference in fixation duration according to the levels of child engagement; 2) there were statistically significant mean differences in fixation duration with pairwise comparisons between all child engagement levels. Combined with the descriptive statistics of fixation features (the average and the total fixation duration), it was further proved that fixation duration contains more information related to children's engagement levels than the overall number of fixations per clip. For example, the average fixation duration increased according to the increased level of engagement. Thus, fixation duration can be used to identify whether a child was engaged or not in the next section.

4.5.2 Classification

A binary classifier (not+rarely engaged vs. highly+fully engaged) was built to classify children's engagement level by taking into account the fixation durations to detect whether a child was engaged in the digital story-stem vignettes. Firstly, a baseline classifier was built that assigned one class which was the most common label. The most common label was the class 'High Engagement' in the dataset. The accuracy of the baseline classifier was 75.86% which means the portion of class 'High Engagement' in the test is 75.86%.

The accuracy of child engagement identification was firstly calculated as it is the most common metric to report classification performance. In the 2-class classification, the accuracy of the classifier was 72.16% (75/104 clips classified successfully). A confusion matrix was firstly calculated as shown in Table 4-13 and it shows that 54 clips were correctly classified as 'High Engagement' and 21 clips were classified as 'Low Engagement'.

Prediction \ Annotation	High	Low
	High	54
Low	10	21

Table 4-13. Confusion matrix of the binary classifier of child engagement (high vs low) using the fixation durations as shown in the number of classified clips.

Traditionally, accuracy was the most common metric to report classification performance. However, the ground truth in this dataset was unbalanced accordingly the baseline: the amount of “High Engagement” class data was greater than the “Low Engagement” class data. Therefore, the classifier that predicted the most frequent class had a deceptively high accuracy. To prevent this problem, various accuracy metrics (see Table 3-9) have been used to evaluate the child engagement classifier using the confusion matrix. The accuracy metrics (see Table 4-14) shows that the classifier using fixation duration measured child engagement correctly in 78.83% of instances with an F₁ score as well as 68.44% of instances with a balanced accuracy, a good performance for this classification task. Moreover, the classifier using the fixation duration has a better performance of completeness than exactness, which means that more clips labelled as high engagement were selected.

Method	Acc.	BAcc.	Prec.	Spec.	Sens.	F ₁ score	MCC
Fixations	0.7216	0.6844	0.7397	0.5250	0.8438	0.7883	0.3922

Table 4-14. Accuracy metrics of the binary classification of child engagement using fixation durations.

4.6 Results of Facial AUs Recognition

The results show the automatic recognition of child engagement levels using their facial action units was possible. Since facial AUs has been used to measure children’s engagement in the context of problem-solving [50], no statistical analysis needed to be presented here to detect if facial AUs contains information related to children’s engagement.

Twenty child participants performed a total of 80 story-stem trials (4 story-stems for each child, S1~S4). The video recordings of 4 story-stems of each child were split into static frames (180s recordings for each child and split into 5400 frames). Aggregating data was performed to recognise the facial action units and recorded in the form of numbers, as introduced in Section 4.3.2. Figure 4-9 shows the extracted facial features of a subset of Action Units (AUs) in each frame and this data was taken from a participant that has an ID as P1 when watching the first story-stem (S1). Each row of aggregated data contained two

types of AU detection: intensity (shown as AU_r) - how intense is the AU (minimal to maximal) on a 5-point scale; presence (shown as AU_c) - if the AU is visible in the face [9].

Figure 4-9. The screenshots of aggregated data. A subset of facial Action Units (AUs) was recognised by intensity (AU_r) and presence (AU_c) and combined with ratings of child engagement levels from human annotation.

For the column titled ‘engagement’ in Figure 4-9, the level of engagement was taken from the annotation results of video recordings for all children. Each frame was annotated with an engagement level $l \in \{1, 2, 3, 4\}$ by two independent raters. However, video recording data were not appropriately recorded for 4018 frames due to bad quality and eye/face occlusion as well as 27422 frames due to inconsistent rating scores from human annotation results. In total, there were 76560 frames and the number (percentage) of frames in each level of child engagement was 5000 (6.53%), 11882 (15.52%), 35501 (46.37%), and 24177 (31.58%) from level 1 to 4 respectively (see Table 4-7).

Therefore, the data used here was a set of the level-of-child-engagement/ a-subset-of-facial-action-units pairs in each static frame. The pair of data was recorded in the forms of numbers. In the following, the ability to detect whether a child was engaged in the digital story-stems using facial action units was firstly evaluated. Then the five most frequent facial action units are used to analyse children’s mood states, such as surprise and fear, based on the classification results.

4.6.1 Classification

Similar to the classification task using eye-tracking measures, a baseline of the 2-class classifier (the AU classifier) was built that assigned one class which was the most common label. The most common label was the class ‘High Engagement’. The accuracy of the baseline classifier was 77.95% which means the portion of class ‘High Engagement’ in the test is 77.95%.

The accuracy of automatic child engagement measurement was firstly calculated as it is the most common metric to report classification performance. In the 2-class classification, the accuracy of the AU classifier was 79.97% (18368/22968 frames classified successfully). In addition, facial AUs can be described in two ways: presence - if the AU is visible in the face;

intensity - how intense is the AU (minimal to maximal) on a 5-point scale [9]. Researchers have trained the intensity and presence indicators separately on different datasets and they found that the predictions of both might not always be consistent [8]. For example, the AU presence model could be classifying the AU as not being present, but the intensity model could be classifying its value above 1. Therefore, another two classifiers were constructed: 1) the AU intensity classifier, using a set of pairs (child engagement level/a set of AU intensity) data; and 2) the AU presence classifier, using a set of pairs (child engagement level/a set of AU presence) data. The accuracy of AU intensity and AU presence classifier was also computed separately.

The accuracy of the AU intensity classifier was 80.54% (18498/22968 frames classified successfully) while the accuracy of the AU presence classifier was 79.03% (18247/22968 frames classified successfully). The accuracy of classifying child engagement using the three factors related to facial AU (i.e., AU intensity, AU presence and a combination of these two factors) shows that the AU intensity classifier had highest rate of classification, which means that the AU intensity classifier had a better performance than the two other classifiers to measure if children were engaged during watching the digital story-stems. Like the classifier using the fixation durations, this facial AU classifier that classified the most frequent class had a deceptively high accuracy because the amount of “High Engagement” class data was greater than the “Low Engagement” class data accordingly the baseline. To prevent this problem, various accuracy metrics were computed including the metrics used in Chapter 3 (see Table 3-9) and the Receiver Operating Characteristic (ROC) curves and its areas under the curve (AUC) [36, 48]. In a ROC curve, the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two classes.

A confusion matrix was firstly calculated for the three classifiers related to facial AUs as shown in Table 4-15 and Table 4-16. Compared to the three confusion matrices, the AU classifier and AU intensity classifier had the same performance as the number of correctly identifying frames labelled as high engagement was the same (16368/22968) while AU intensity classifier had a better performance of classifying low engagement levels than AU. Although the AU presence classifier identified high engagement frames less accurately, it worked better on classifying low engagement levels than classifiers using AU and AU

intensity values. Accuracy metrics were calculated using the three confusion matrices. Table 4-17 shows that all three classifiers had a high precision and sensitivity values. This means that the three classifiers had good performance of identifying frames as high engagement levels correctly as well as the frames that correctly identified as high engagement levels accounted for a large proportion (94% for the AU and AU intensity classifier and 92% for the AU presence classifier) in all correctly identified frames. The accuracy metrics also shows that the AU intensity classifier had the best accuracy in F₁ score and MCC while the classifier using AU presence had the lowest accurate rate. The classifier using facial AUs is the combination of the data taken from AU intensity and AU presence, therefore its performance had a lower accuracy than the classifier using AU intensity.

Prediction \ Annotation	High	Low
	High	16368
Low	929	2000

Table 4-15. Confusion matrix of the binary classifier for child engagement using facial AUs.

Prediction \ Annotation	High	Low
	High	16368
Low	928	2130

Prediction \ Annotation	High	Low
	High	15980
Low	1321	2267

Table 4-16. Confusion matrices of the binary classifier for child engagement (high vs low) using the facial AU-intensity (left) and using AU-presence (right).

Method	Acc.	BAcc.	Prec.	Spec.	Sens.	F ₁ score	MCC	AUC
Facial AUs	0.7997	0.6495	0.8168	0.3527	0.9463	0.8768	0.3865	0.8377
AU intensity	0.8054	0.6609	0.8221	0.3755	0.9463	0.8799	0.4086	0.7801
AU presence	0.7945	0.6618	0.8246	0.4000	0.9236	0.8713	0.3843	0.7760

Table 4-17. Accuracy metrics of the three classifiers for child engagement (high vs low) using the facial AU, AU intensity and AU presence respectively.

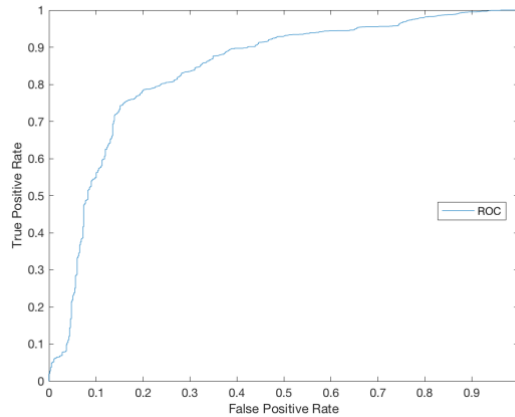


Figure 4-10. ROC curve of the binary classification for child engagement (high vs low) using the facial AU.

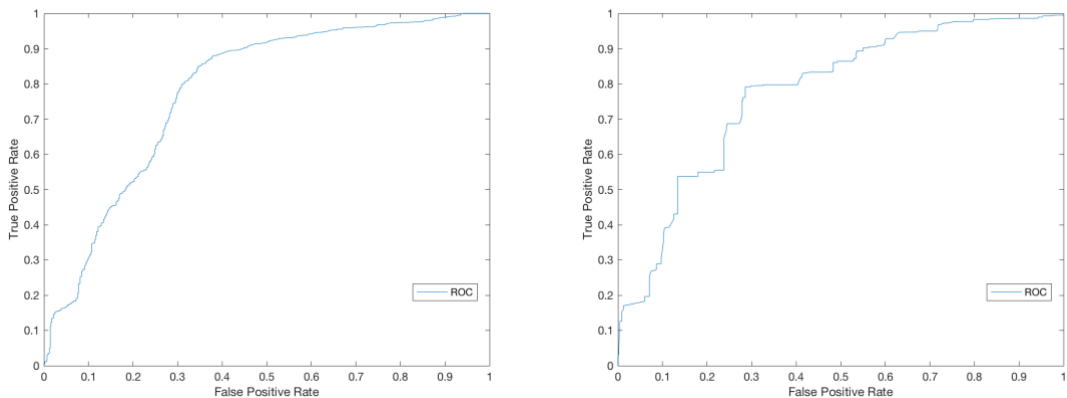


Figure 4-11. ROC curve of the binary classification for child engagement (high vs low) using facial AU-intensity (left) and AU-presence (right).

The ROC curve is a performance measurement for classification problem at various thresholds settings, which was shown in Figure 4-10 for the AU classifier and in Figure 4-11 for the AU intensity classifier (left) and the AU presence classifier (right). AUC tells how much model is capable of distinguishing between classes. The AU classifier had the highest value of AUC with 0.8377 while the lowest AUC value was 0.7760, calculated from the AU presence classifier. It indicates there was 83.77% of chances that the AU intensity classifier would be able to distinguish between high engagement class and low engagement class while the AU presence classifier had 77% of chances. The results from the ROC curve differed from other various metrics such as F1 score and MCC, which means that although AU presence showed poor performance when used independently, it provided additional discriminative information while used in combination with AU intensity. Considering the ability of a classifier to detect whether a child is engaged in the digital story-stem vignettes, the result suggested that children's engagement (high versus low) could be detected using the facial AU intensity classifier.

4.6.2 The Most Frequent AUs of Engagement

A descriptive analysis was conducted into how facial action units performed when children were engaged based on the classification results using the feature of AU intensity. For example, frames that classified as low engagement always had a high intensity of AU45 (eye closure), while frames were classified as high engagement by exhibiting high intensity of AU2 (outer brow raise). The most five frequent facial action units by calculating its intensity from frames that classified with high engagement (17296 frames in total) was: AU01 (inner brow raise), AU02 (outer brow raise), AU12 (lip corner pull), AU14 (dimple) and AU17 (chin raiser). Table 4-18 shows that the number of frames that contained each frequent AU intensity (the intensity value > 0) and the proportion was the number of frames with each frequent AU intensity divided by all high engagement frames. The number of frames with corresponding presence (the presence value = 1) of these five action units was also calculated. The remaining facial action units occurred less frequently.

Facial Action Unit	Intensity	Presence
AU01: Inner Brow Raise	7664 (44.31%)	4727 (27.33%)
AU02: Outer Brow Raise	10214 (59.05%)	6477 (37.45%)
AU12: Lip Corner Pull	5068 (29.30%)	7835 (24.49%)
AU14: Dimple	11550 (67.78%)	10808 (62.49%)
AU17: Chin Raiser	6142 (35.51%)	4101 (23.71%)

Table 4-18. The five most frequent facial action units (AUs) intensity and presence from the classification result, as shown in the number of frames and in parenthesis in per cents.

Frames with high intensity of AU02 and AU14 are typically classified into high engagement categories. The prototypical “fear brow” combines AU01, AU02, and AU04 and it is a highly reliable indicator of “fear” because it is so “difficult to make deliberately” [24, 35]. Researchers have revealed feelings of anxiety facially corresponded to the elements of the expression of fear [35]. The primary fear elements were the eyebrow actions and horizontal mouth stretch movement (AU14). Based on the context of the MCAST test, the fear elements could be explained that a child was engaging in the story with a situation of specific anxiety and distress while fearful facial actions were displayed, i.e., AU02 and AU14 occur in frames as primary actions. Furthermore, almost half frames (44.31%) contained the intensity value of AU01 while the number of frames with AU01 presence were much less than its intensity. It means that AU01 was a frequent action unit but there was not an obvious eyebrow movement. Therefore, AU01 was hard to observe by human observers and frames were hard to be predicted only based on the AU01.

4.7 Results of the Self-report Measure

Each child was asked to fill in a questionnaire after completing each MCAST story-stem vignette. Their answers were used to give a support for better understanding of children's engagement levels. In this questionnaire, Q8 and Q9 are open-ended questions about the feelings of the child doll and the mummy doll respectively that related to story understanding. Seven single choice questions (Q1- Q7, see Table 4-8) were using a smiley-face based 5-point Likert scale to investigate four aspects of child engagement. Children's answers of choosing a smiley face for each question were transformed into the form of numbers (between 1~5), "Not at all true" using a totally sad face was coded as one and "really true" using a totally happy face was coded as five. A descriptive statistical analysis (see Table 4-19) of children's answers for four aspects of the questionnaire shows that children think the story-stems are easy to understand and they pay attention to them.

Aspects	Related Questions	Number of children	Mean	S.D.
Distraction/attentional focus	Q1, Q2, Q4	20	3.93	1.142
Story understanding	Q3	20	4.12	0.805
Empathy	Q5, Q6	20	3.13	1.531
General attitudes/ Interest	Q7	20	3.68	1.430

Table 4-19. Descriptive analysis of children's answers for the questionnaire according to the four aspects of child engagement measurements.

The aspect of distraction/attentional focus aims to measure children's concentration and absorption in each story. The results of children's answers to questions related to this aspect indicated children were able to pay attention to the story as well as were not distracted by other things (3.93/5). The average rating score of Q2 was 3.85/5 (S.D. = 1.261) that indicated children were happy to complete the MCAST story vignette. The completed story and children's behaviours would be used for attachment assessment. If a child could complete the story spontaneously, MCAST assessors or the SAM system could collect more reliable data to evaluate the child's attachment status.

For the aspect of story understanding taken from the questionnaire, children's answers show that all MCAST story-stems were easy to understand and about 60% of children recognised stories were quite easy to understand.

The aspect of empathy focuses on measuring if child can feel with the child doll's emotions represented as distress due to a predicament shown in each story-stem. Children's answers

showed that they could feel with the child doll's emotion (average ratings 3.79/5 for Q6) but they did not think they have a stressed feeling during following of the story video on screen (average ratings 2.47/5 for Q5). In addition, children's answers of Q8, the open question to ask the child doll's feelings, indicated that they can feel with the child doll's 'bad' emotions due to a "bad" situation for the child doll represented in the story-stem, and child participants think the child doll feels better in their completed stories. For example, an answer of Q8 was "sad with a tummy ache but happy now" from the illness vignette.

Overall, children think this task was interesting (3.68/5 for Q7). 40% of children chose "Yes, I really like them!" (coded as 5) for the MCAST stories. For children who dislike the MCAST stories (rating less than 3), 15% of children chose "No, I don't like them at all!" (coded as 1) and only 5% chose "No, I don't like them." (coded as 2). This indicated that children had a strong attitude to express their dislike.

4.8 Discussion

The problem of child engagement evaluation in various contexts has generated great interest in methods to measure it. One important area, the story-stem approach, has been recognised as a reliable and cost-effective method for assessing child Attachment status. Due to high cost and time required for conducting the assessments, a computer-based tool is being developed for automate attachment assessments in a cost-effective way. However, providing such tests via computer relies on the child being engaged in the story. The instance of the story-stem approach used here was the Manchester Child Attachment Story Task (MCAST). This chapter proposed a method of child engagement level measurement in the context of digital MCAST story-stem viewing.

Facial expressions were collected from 20 children using an RGB webcam as well as a desktop eye-tracker while they watched the story-stems from MCAST to investigate whether children's spontaneous facial expressions can be used to automatically measure their engagement levels in digital story-stems (RQ1). Two methods based on computer vision provided an automatic estimation of engagement by analysing cues from the facial muscles and eyes respectively. In addition, children were also asked to fill in a questionnaire and their answers could give a better understanding of their engaged states.

The analysis procedure for child engagement recognition using the eye-tracking technique was the same as in Chapter 3. Fixation is the main measurement and two fixation metrics

(the number of fixations per clip, fixation duration) were analysed whether they contained information related to different child engagement levels in digital MCAST story-stems. Descriptive statistics firstly show the distribution of the two fixation metrics according to child engagement levels. The total number of fixations and the total fixation duration has the same distribution as the distribution of clips in terms of child engagement. The highly-engaged category (level = 3) has the highest total number of fixations as well as the longest total fixation duration as clips labelled as level 3 are the most frequently occurring. Meanwhile, when child engagement levels increased from level 1 to 4, the average number of fixations per clip decreased (not much difference between level 3 and 4) while the average fixation duration per clip increased.

Two one-way ANOVAs using two fixation metrics separately, and their *post hoc* tests, show that there were statistically significant differences in fixation duration across the four child engagement levels, as well as there were significant pairwise mean differences in fixation duration between all child engagement levels. Although there were also significant differences in the number of fixations per clip across the four child engagement levels, the result of the *post hoc* test shows the mean differences in the number of fixations per clip with pairwise comparisons between all child engagement levels, were not always statistically significant. Therefore, the statistical result showed that fixation duration was a better indicator than the overall number of fixations per clip for engagement measurement and fixation duration can be used to classify children's engagement. The classification task shows that the SVM classifier using fixation duration measured child engagement correctly in 78% of instances with an F_1 score, a good result for automatic engagement recognition.

Moreover, combined with Chapter 3, fixation duration had significant differences across the four levels for both adult and child engagement. The descriptive statistics shows that the mean fixation duration per clip for both adults and children increased when the engagement level increased from level 1 to 4. Compared to the adult group, children's mean fixation duration per clip was longer than the adult group across the engagement levels. Longer fixation durations mean that participant's gaze behaviour changes from a slow-paced information extraction to a higher comprehension in story-stems viewing on screen. Meanwhile, the descriptive statistics of the average number of fixations per clip shows that it decreased according to an increased level of engagement for both the adult and child group and children have a higher number of fixations per clip than adults across the four engagement levels. Compared to the number of fixations per clip in each engagement level,

there was not much difference in the number of fixations per clip between level 1 and 2 for the adult group while it was similar between level 3 and 4 for the child group. The statistical analysis indicated that the number of fixations per clip is not a good indicator as the mean differences in it with pairwise comparisons between all engagement levels of both adult and child, were not always statistically significant. Thus, fixation duration is a good indicator for engagement classification for both the adult and child group.

The classification task shows that gaze behaviours contain information related to levels of child engagement. The SVM classifiers using fixation duration measured both adult and child engagement correctly in about 68% of clips with a balanced accuracy (67.27% for adults and 68.44% children), a good result for automatic engagement identification. A balanced accuracy was then calculated by the average of the sensitivity value (i.e., the percentage of clips with high engagement that are correctly identified) and the specificity value (i.e., the percentage of clips with low engagement that are correctly identified). Sensitivity and Specificity are inversely proportional to each other. When the sensitivity value increases, the specificity value decreases and *vice versa*. Compared to the sensitivity and specificity values of the classification performance between the adult and child group, the classifier for child engagement had a higher sensitivity value and a lower specificity value than the classifier for adult engagement in Chapter 3. For example, the specificity value was 0.5250, which means 52.50% of clips was correctly identified using children's fixation durations. It was much lower than the specificity value using adult fixation durations, correctly in 71.05% of clips. This demonstrated that more clips labelled as high engagement were identified correctly and fewer clips labelled as low engagement identified correctly. An F_1 score was the harmonic mean of precision and sensitivity, both the precision and sensitivity values were related to the exactness and completeness of the classifier for identifying clips with high engagement. While analysing the classification performance using an F_1 score, the classifier measured child engagement correctly in 78% of clips, higher than the adult engagement classification (70% of clips) and both precision and sensitivity values were higher in the child group than the values in the adult group. This was caused by the number (percentage) of high engagement clips in the child group was much higher than the number (percentage) in the adult group (75.86% vs 59.24% of high engagement class). The ability of the classifier was affected by the distribution of classes. Moreover, due to the unbalanced class distribution and low accuracy of low engagement identification, this dataset would not be used to build a 4-class classification task for child engagement as the accuracy of classifying clips with low engagement may be further reduced.

Besides using eye-tracking, the video recordings were spilt into static images and facial action units were extracted for every frame using OpenFace. Human labellers were asked to manually annotate perceived child engagement levels based on the annotation scheme and an automated system was built to identify the engagement levels of children from their facial action units. Based on human annotation and selection results in terms of child engagement, the actual level of child engagement was successfully recognised as a binary classification using a set of child-engagement/ facial-action-units pairs, in which not and rarely engaged levels were grouped into a 'low engagement' class, and highly and fully engaged categories were grouped into a 'high engagement' class.

The results of the classifier show that facial AUs contained information correlated with children's engagement levels. The accuracy of three classifiers using different factors related to facial AU (i.e., the AU intensity classifier, the AU presence classifier and a combination of these two factors called the AU classifier) shows that the highest accuracy rate of classification was the AU intensity classifier. Due to unbalanced distribution of child engagement (the amount of "high engagement" class data was greater the "low engagement" class data), accuracy metrics were calculated to test the performance of the three classifiers. The accuracy metrics demonstrated that the classifier using AU intensity had a better performance in identifying levels of child engagement than two other AU-related classifiers (correctly in about 87% of cases with an F_1 score). The best subset of facial Action Units was then analysed. It included facial movements of eyebrow, mouth and chin. Based on the context of the MCAST test, the frequent action units could explain that a child was engaging in a story-stem with a situation of specific anxiety and distress, while fearful facial actions were displayed, i.e., AU02 and AU14 occurred in frames as primary actions. Furthermore, frames were hard to identify based only on the AU01 because almost half frames (44.31%) contained the intensity value of AU01 while the number of frames with AU01 presence were much less than its intensity.

Additionally, this dataset would not be used for a 4-class classification task in terms of child engagement due to the unbalanced distribution of classes (77.95% of frames labelled as high engagement). The specificity values, the percentage of the accuracy of identifying frames with low engagement levels using three AU-related classifiers were in 37.55%, 40.00% and 35.27% of frames respectively, a poor performance for identifying low engagement. Thus, the accuracy of a 4-class classification task for child engagement may be further reduced.

Thus, compared to the two measurement methods (eye-tracking and facial AU recognition) used in this chapter, facial AU recognition had a better performance for identifying engagement or disengagement for children than the eye-tracking measure based on the results of the accuracy metrics. The accuracy metrics show that the sensitivity and precision values for the AU intensity classifier, the best classifier among different AU- related classifiers, were higher than the classifier using fixation duration, which means a better exactness and completeness for identifying the ‘engaged’ children.

The self-report data were analysed whether children’s answers of the questionnaires was the same as their engagement levels identified using the automatic methods for their spontaneous facial expressions. Overall, the descriptive statistics of the questionnaire shows that children think they were able to understand and pay attention to watching the story-stems. This was consistent with the annotation results of children’s facial behaviours from the video recordings, which child participants had an overall high engagement levels during following of the story-stems video on screen. The eye-tracking measurement method used children’s fixation durations to measure if a child could understand the story-stem during following of the story-stem videos on screen. The sum of fixation durations indicated that engaged children have longer fixation durations (more than 50% of watching time) than disengaged children when watching the story-stems video on screen. The literature indicated that longer fixation durations means a higher comprehension as participant’s gaze behaviour changes in a slow pace for extracting information. Also, children’s answers of the questionnaire (Q3) show that they think the MCAST story-stems are easy to understand. Thus, the eye-tracking measure was a good measurement method of story understand, an important aspect of child engagement in the context of digital story-stem viewing.

While analysing the aspect of empathy taken from the questionnaire, children’s facial muscle movements (recorded using facial action units) could reflect their mood states. The MCAST stories aim at increasing the child’s mood state around the particular distress represented in the story. Children may have a general increase of emotion and feel pressure when watching the stories, which will have activated their mental representation of attachment. The analysis of the frequent facial action units shows that children often had “fearful” facial actions, which can be used to explain that a child was engaging in the story with a situation of specific anxiety and distress. However, children’s answers related to the aspect of empathy in the questionnaire (Q5 and Q6) show that they can well understand the child doll’s emotion while they did not think they have a stressed feeling during following of the story video on screen.

Their answers were incompatible with the analysis of facial action units. Although children's answers said they do not think they feel stressed, the analysis of their facial expressions already catch their "fearful" faces. This means that children's answers of the questionnaire for their mood state were not always the same as the performance of their facial expressions, which could be collected and analysed using automated methods. The self-report measures for children may not be the most suitable method for measuring their engagement levels. Children's spontaneous facial expressions recorded by a webcam could be used to analyse and reflect their mood state, but child participants may not express their attitudes and emotions accurately towards the study by filling in a smiley-o-meter questionnaire.

4.9 Conclusion

This chapter focuses on measuring child engagement levels in digital story-stems using their spontaneous facial expressions. The main contribution of this chapter was to analyse gaze behaviours and facial action units for automated child engagement recognition to identify 'engaged' children in the context of digital stories viewing, which can be implemented with low cost algorithms and in a non-invasive way with simple sensors, which was answered RQ1. Also, a comparison of results for measuring child (in this chapter) and adult (in Chapter 3) engagement levels using gaze behaviours shows that fixation duration can be used to measure the engagement levels across different age groups.

The analysis of children's spontaneous facial expressions included eye-tracking measures and facial AU recognition, both methods contained information related to the distinct levels of child engagement. The facial AU recognition had a better performance for identifying 'engaged' children than the eye-tracking measure, according to the classification accuracy with various metrics. Large amounts of spontaneous facial actions could be acquired in order to explore the causes and variables that affect child engagement. Future work will focus on collecting more data for the not engaged and rarely engaged categories to appropriately address the 4-class classification problem.

The instance of the story-stem approach used in this thesis is the Manchester Child Attachment Story Task (MCAST). There was a high-level of engagement of children in this study, which suggested that digital story-stems could be used for child psychiatric studies. The contribution of automated child engagement measurement reduces the need for so much time from trained administrators and ensures the quality of the data that will be used to make

assessments, improving the efficiency of coding Attachment evaluations. People without MCAST training, such as teachers, could administer the MCAST test to reduce the cost and involvement of fully trained administrators, who could then reserve their time for assessment of the video data to make attachment ratings and treat the children. Automating this process could also be useful for other psychological and educational studies which use the same type of protocol.

Chapter 5 Designing an Engaging Digital Story-stem

5.1 Introduction

The story-stem approach with traditional storytelling is a reliable and valid assessment method for investigating the important relationships in a child's life, and has made significant contributions to Attachment theory [14, 15, 32, 77]. Engagement is an important concept in the tests using story-stems, where it is in initial phase of the test – a child is given the beginning of a story by an assessor using two dolls. Bringing the child into a deep engagement with a story is a key step to bring out his/her mental representation of attachment. Engagement means that children focus on attending to the play and materials, are not distracted by other things, and feel empathy with the dolls and characters in the story. If a child is not emotionally engaged by the predicament shown in each story-stem, the psychiatrist cannot assess their attachment status based on their story and behaviour during the activity.

The instance of the story-stem approach used in this thesis was the Manchester Child Attachment Story Task (MCAST). Since the use of the story-stem approach in the MCAST test takes a lot of time and there are often few administrators trained to administer them, a system called SAM [78, 87] is being developed that presents the digital story-stems on a laptop screen to successfully digitalise the interaction between the child and the story without disrupting the storytelling.

With the arrival of multimedia, digital story-stems can be constructed using a mixture of graphics, animation, text, recorded audio narration, video and music, to present information on a specific topic [76]. SAM, as discussed in Section 2.3, uses live-action videos for displaying the MCAST stories. In order to bring a child into a deep engagement while the child was watching the digital story-stems, the focus in this chapter is on designing an engaging MCAST digital story. Two key aspects to be studied are: the storytelling voice and the video format. For example, a good voice in a digital story makes audiences fit the story

line and really “get into” the story while audience may not be consistently engaged in a flat voice which does not fit the story line [86]. One purpose of this thesis is to apply the story-stem approach with multimedia tools and detect whether a child is engaged with the digital story-stems. So, do these multimedia types affect child engagement levels in digital story-stems?

Chapter 4 described a method for measuring child engagement levels using spontaneous facial behaviours while children were watching digital story-stem vignettes. During that study, the story-stems were displayed using short live-action recorded videos. The live-action videos used was based on the SAM videos, where involves the video recordings of an adult storyteller using two physical dolls (a mummy doll and a child doll) and playing with a real dolls-house to show each MCAST vignette. However, this may not be the best way of engaging the children, as other multimedia types of presentation are available. Compared to live-action videos, animation as another video format for the presentation type has been used for child education and the use of animations can attract children's attention to a certain part of the screen during the storytelling processing [37]. These studies suggest that animation might be a good alternative for the design of the MCAST story-stem vignettes in this thesis. Both of these are possible to use in story-stem vignettes and this chapter will investigate which is most effective. Meanwhile, as an emotional storytelling voice could capture audience's attention to create an engaging listening experience, the storytelling voice will also be studied for the effect on children's engagement levels.

This chapter investigates the role of storytelling voice and video format (animation vs. live-action video) as two multimedia types for engaging children in digital MCAST story-stems respectively. The first part of this study focuses on investigating the role of storytelling voice on engaging children in MCAST stories. Two storytellers (one male and one female) were asked to record each MCAST story-stem in two ways with different expressive qualities (expressive voice vs flat voice). The second part investigates the effect of animated MCAST videos on child engagement levels and compares it to the live-action videos. The animated story-stems display movement of simple two-dimensional symbolic screen ‘dolls’ while the live-action story video from Chapter 4 display movement of two physical dolls used by an interviewer. Both presentation types were narrated by the same audio, taken from a human storyteller's voice with different voice conditions.

Therefore, this chapter aims to answer the RQ2: *How do voice type and presentation type affect child engagement levels in digital story-stems?* Facial data from 40 children were recorded by an RGB video camera while watching the MCAST story-stems. A statistical analysis of the engagement levels was conducted across different multimedia types. By analysing different multimedia types of a digital story and children's story experience, this chapter is to gain a better understanding of how to produce an engaging story to children using different multimedia technologies.

5.2 Evaluating the effects of media type on engagement

The MCAST story-stem vignettes were redesigned using animation tools and represented on the computer by the movement of two symbolic screen 'dolls', narrated by two storytellers with flat and expressive voices.

5.2.1 Storytelling Voice

The MCAST story-stems were recorded with two between-subjects voice conditions: voice gender (female vs male voice) and voice expressiveness (expressive vs flat voice). To control the differences, e.g., pronunciation and quality, one female and one male adult storyteller recorded both expressiveness types (the expressive and the flat voice) for each story-stem. For the Expressive condition, the utterances were emotive with a larger dynamic range; the storytellers were instructed to speak in an expressive, emotional way. For the Flat condition, storytellers imitated a text-to-voice voice, keeping their intonation very flat. Computer-generated voices were not used for telling the stories because it is hard to make them expressive enough; no computer-generated voices can currently imitate the dynamic, expressive range of human storytellers' voices. Therefore, we used actors, recruited to create the voices needed.

There are four types of the storytelling voices in this study: female expressive (FE), female flat (FF), male expressive (ME) and male flat (MF) voice. A check of the voices using the pitch and intensity was performed to ensure recordings with different voice conditions were actually different, specifically on whether the expressive voice recordings would be perceived as more emotional and expressive than the flat voice recordings. From the literature review, the proper value of storytelling voice is taken from converting synthetic speech to expressive speech and it did not mention voice gender. So, it is assumed that the

increase between 30Hz and 60Hz from the flat to expressive voice is reasonable. The best value provides a range rather than a perfect value.

The fundamental frequency F0 (also called pitch) is a variable that distinguishes differences between the voices of male and female adult speakers. In general, adult males tend to have voices with a low F0 or low pitch, and adult females tend to have voices with a high F0 or high pitch [69]. The audio files were exported from the video recordings and Table 5-1 shows the key variables of pitch values including a maximum, minimum, mean, and standard deviation using the PRAAT²⁴ system. A one-way ANOVA was performed to test for an effect of different storytelling voice types on pitch values. The data used here was a set of pitch-value/voice-type pairs recorded in the forms of numbers. The test modelled the differences in the mean of the pitch values as a function of type of the storytelling voice types. It indicates a statistically significant difference in pitch values according to the four storytelling voice types ($F(3, 4399) = 240.311, p < .001$).

PITCH (Hz)	Voice conditions			
	FF	FE	MF	ME
Max	739.950	764.650	687.978	714.307
Min	113.144	153.805	99.780	102.600
Mean	235.572	331.105	218.902	226.372
SD	82.060	91.456	163.531	130.876

Table 5-1. The pitch value of each storytelling voice type (FF = female flat, FE = female expressive, MF = male flat, ME = male expressive).

Furthermore, to investigate where the actual differences are in the ANOVA test, a Hochberg²⁰ *post hoc* test was conducted to list pairwise differences among groups for the independent variable (the storytelling voice types). The results of the *post hoc* test (see Table 5-2) shows the differences between the mean number of fixations in each clip, the p-value and its standard error for multiple pairwise comparisons under a 95% confidence interval. There were significant pairwise mean differences in pitch values between the female expressive voice type and the other three voice types separately, with an average difference of 95.422 ($p = .000$) between type FE and FF; with an average difference 112.211 ($p = .000$) between type FE and MF; and with an average difference of 104.622 ($p = .000$) between FE and ME. It indicates a significant difference in pitch values between the female expressive voice type and the other three types respectively.

²⁴ <http://www.fon.hum.uva.nl/praat/>

However, the differences in pitch values between other storytelling voice types were not always statistically significant. For example, ME has a significant difference in pitch values with type FE, with an average difference -104.622 ($p = .000$), but the differences in pitch values between MF and other storytelling voice types, was not statistically significant. This means the male storyteller narrated the story-stem in his flat and expressive voice, the mean pitch values were similar. Therefore, another check was then performed to ensure the expressive recordings were sufficiently more emotional and expressive than the flat recordings under the same voice gender, so that storytelling voices could be used to investigate which type would be the best voice for creating an engaging digital story-stems.

Dependent Variable: Pitch value				
Voice type v	Voice type v'	Mean Difference ($v-v'$)	Std. Error	p-value
FF	FE	-95.422*	4.724	.000
	MF	16.974*	5.352	.010
	ME	9.274	4.881	.308
FE	FF	95.422*	4.724	.000
	MF	112.221*	5.240	.000
	ME	104.622*	4.759	.000
MF	FF	-16.974*	5.352	.010
	FE	-112.221*	5.240	.000
	ME	-7.596	5.383	.644
ME	FF	-9.274	4.881	.308
	FE	-104.622*	4.759	.000
	MF	7.596	5.383	.644

Table 5-2. Result of a Hochberg *post hoc* test to find the actual differences of the four storytelling voice types using the pitch values. *: The mean difference is significant at the 0.05 level.

In the check of the expressive storytelling style, sudden climax and increasing climax as two types were used to express suspense in the stories. Sentences that contain the two types was used to check whether the expressive recordings were sufficiently more emotional and expressive than the flat recordings under different voice genders respectively. As each MCAST story-stem contains a 'distress' situation, the 'distress' situation in the story was recognised as a sudden climax in each story. The increasing climax was not considered to check the voice expressiveness in the MCAST story-stems because the dramatic event, like the predicament, cannot be expected in advance.

Story-stems	Selected Sentences
Nightmare	Then suddenly the child doll wakes up. And he says “Ooooh... I’ve had a horrible dream ooh... a horrible horrible dream...”
Hurt Knee	It’s almost the end... And Oh no! The child doll slips in a puddle! “Ooooh...” he cries “I hurt my knee... and it’s bleeding ...”
Illness	Suddenly the child doll has a pain in his tummy. And it gets worse! The child doll cries “Ooooh... I’ve got a pain in my tummy...”
Shopping	The child doll feels very scared and he cries “Ooooh...where’s my mummy? Where’s my mummy?”

Table 5-3. The four sentences that contain a sudden moment from the MCAST story-stems to be used to check the differences between the flat and expressive recordings.

Four sentences were selected from the MCAST story-stems as shown in Table 5-3 each sentence contained a sudden moment from each MCAST story-stem. The sudden climax with different voices type was displayed using a pitch contour by a dramatic increase of pitch on the keyword that was located in the dash. Figure 5-1 was shown for the female voices and Figure 5-2 for the male voices. The x-axis represents time length of the selected fragment and the y-axis represents the pitch values.

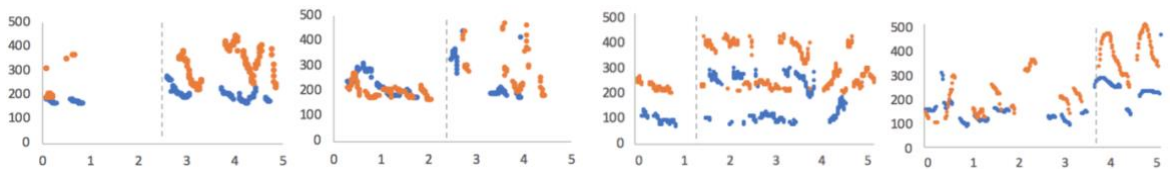


Figure 5-1. The pitch contour of sudden climax for the MCAST story-stems narrated by a female voice (blue = flat voice, orange = expressive voices), from left to right: nightmare, hurt knee, illness, and shopping.

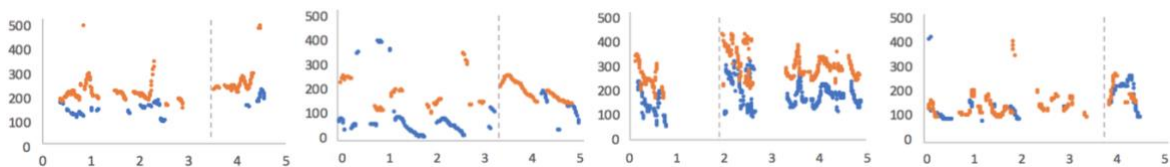


Figure 5-2. The pitch contour of sudden climax for the MCAST story-stems narrated by a male voice (blue = flat voice, orange = expressive voices), from left to right: nightmare, hurt knee, illness, and shopping.

Within the time domain $[t_1, t_2]$ (from the dash to the end of each fragment), it can be seen that the expressive voices from the two storytellers have a larger rise-fall pattern in pitch as compared to the flat voices. For example, in the nightmare vignettes, the pitch is around 200Hz at t_1 and increases to 400Hz at t_2 (the dash) for the expressive condition while a gradual increase of pitch in the female flat voice was found which increases from about 190Hz at t_1 to 300Hz at t_2 (the dash). While in the male voice conditions, the overall pitch values were lower than the pitch in female voices. The pitch in the nightmare vignette is

around 200Hz and increases to 310Hz for the expressive condition while increases from around 200Hz to 260Hz for the flat voice condition.

5.2.2 Live-action Videos vs. Animated Videos

Animated version of the MCAST videos were created. In order to be close to the original MCAST test, the recorded videos show that an adult storyteller was standing behind the dolls-house that played with two physical dolls (a mummy doll and a child doll) to display each MCAST story-stem. For investigating the storytelling voice, two adult storytellers (one female and one male) were asked to present the MCAST story-stems respectively. Their voices were also exported as an audio file to be used for the animation design. The screenshots were shown for the nightmare and illness story-stem respectively (see Figure 5-3).

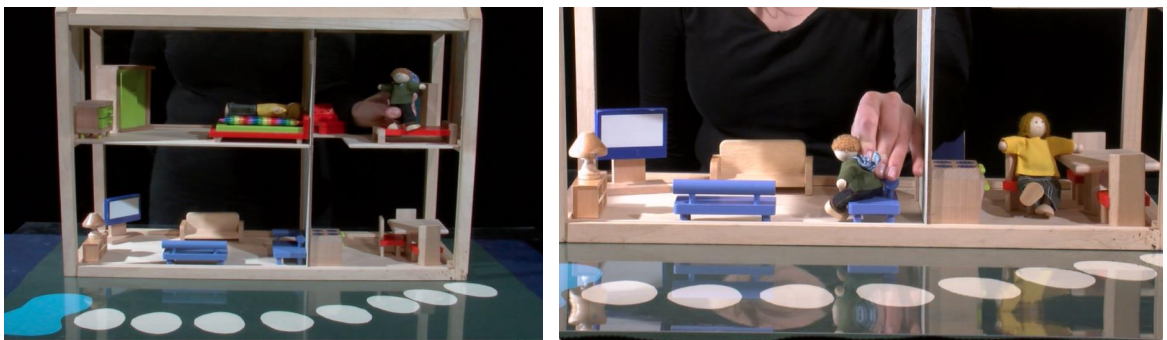


Figure 5-3. A group of screenshots of the MCAST story-stem vignettes displaying as a live-action video. Left: the nightmare story-stem; Right: the illness story-stem.

The animated MCAST videos were made using the CrazyTalk Animator 2 software (CTA 2)²⁵, which enables users to create 2D audio lip-syncing character templates and display movement of characters using motion libraries as well as a bone rig editor. The audio in the animated MCAST stories was the same as it was in the live-action videos. The exported audio files from the live-action videos were imported to the animation. Figure 5-4 shows four screenshots of the MCAST story-stems respectively. For example, the top left one was the ‘nightmare’ vignette, where the child doll wakes up in the middle of the night because of a terrible nightmare while the mummy doll is sleeping in her bed.

²⁵ <https://www.reallusion.com/crazytalk-animator/>



Figure 5-4. A group of screenshots of the MCAST story-stem vignettes displaying as an animated video. (The displayed story in animation was nightmare, hopscotch, illness and shopping from the top left one to the bottom right one respectively.)

There were some differences between the live-action video and animated video. The first difference was specifically focused on the ‘nightmare’ story-stem. A two-layer house with four rooms was displayed in the live-action video while only two bedrooms (one layer) was designed in the animated video. MCAST experts suggested that displaying two bedrooms in animation was acceptable because this story happened only in the bedrooms. Similarly, only displaying the living room and kitchen in the illness story-stem was also acceptable.

Secondly, the characters and furniture between the two video formats were not looking at totally the same. The physical dolls and furniture used in the live-action video was also used as the experimental equipment to help children complete the story-stem for assessing their Attachment status. However, in the animated video, the shape of characters and furniture was not the same as the physical dolls used in the tests. This may cause problems in handing over of initiative to the child that triggers the next phase (story completion), such as cannot distinguish the mommy doll and the child doll. After discussing with MCAST experts, a short animation was designed and displayed to children at the beginning of the test. The short animation gives an introduction of the two symbolic dolls (one mommy and one child) on screen as well as the physical dolls prepared, which could help children distinguish the dolls. Thus, these differences of the two presentation types were accepted and measuring children’s engagement levels in the animated videos that compares it to children’s engagement levels in the live-action videos is the area of interest in this study.

5.3 Methods

5.3.1 Participants

Forty children (7-10 years old, 20 males and 20 females) were recruited from several Glasgow (UK) schools based on their school and parental agreement. After a school agreed to participate, classroom teachers sent opt-out consent forms to each child's family. The forms are shown in Appendix D. This informed families about the research project, explaining the research into a better understanding of child engagement in digital story-stems, introducing the types of data to be collected and the tools to be used for data collection. In the event that a child's family opted out of participating in the research, the child was not selected to participate.

	Story: Nightmare	Story: Hurt Knee	Story: Illness	Story: Shopping
Group 1	Animated Video + Female Flat	Animated Video + Male Flat	Animated Video + Female Expressive	Animated Video + Female Expressive
Group 2	Animated Video + Male Expressive	Animated Video + Female Expressive	Animated Video + Female Flat	Animated Video + Male Flat
Group 3	Animated Video + Male Flat	Animated Video + Female Flat	Animated Video + Male Expressive	Animated Video + Male Expressive
Group 4	Animated Video + Female Expressive	Animated Video + Male Expressive	Animated Video + Male Flat	Animated Video + Female Flat
Group 5	Recorded Video + Female Flat	Recorded Video + Male Flat	Recorded Video + Female Expressive	Recorded Video + Female Expressive
Group 6	Recorded Video + Male Expressive	Recorded Video + Female Expressive	Recorded Video + Female Flat	Recorded Video + Male Flat
Group 7	Recorded Video + Male Flat	Recorded Video + Female Flat	Recorded Video + Male Expressive	Recorded Video + Male Expressive
Group 8	Recorded Video + Female Expressive	Recorded Video + Male Expressive	Recorded Video + Male Flat	Recorded Video + Female Flat

Table 5-4. The allocation of children to watch the MCAST stories with different media types.

To investigate the effect of different media types on engaging children in digital story-stems, the 40 children were divided into two groups. Each group had 20 children (10 males and 10 females), one for watching the animated MCAST videos and another for watching the live-action MCAST videos. Then each group were divided into four small groups for different

voice conditions. In total, there were 8 groups and each group had 5 children. The displayed MCAST stories with media types were allocated to each group as shown in Table 5-4.

5.3.2 Procedure

The test took approximately 20 minutes for each child. The procedure of this study was the same as the procedure from Chapter 4 (see Section 4.2.2). The smiley-o-meter questionnaire was shown in Appendix F about their story experience and opinions to the media types after they watched the digital story-stem.

5.3.3 Data Annotation and Selection

After data collection, the next step was to label the video recordings in terms of engagement. The annotation scale is taken from Chapter 4 (see Table 4-1). Chapter 4 provided automated engagement measures that extracted two kinds of facial features: facial Action Units (AUs) from static frames and gaze behaviours short video clips respectively. Although both facial features successfully measured children's engagement in Chapter 4, the performance of accuracy metrics shows that facial AU recognition had a better performance for identifying engaged or disengaged for children than the eye-tracking technique. Also, human labellers (same as Chapter 3 and 4) indicated that annotating perceived child engagement levels in short video clips manually was more difficult than labelling frames as the engagement level may vary across a clip. Also, the performance of two classification tasks showed that facial action units contained more information related to child engagement levels than gaze behaviours due to the high accuracy of the classifier. Therefore, all recordings were split into static and each one second video clip was split into 30 frames. The procedure of data annotation and selection of this study was the same as the procedure from Chapter 4 (see Sections 4.2.3 and 4.2.5).

5.4 Recognition of Child Engagement

There were 216000 frames (120mins of recordings) in total, without considering different multimedia conditions. The agreement among human annotators were then calculated. Since the engagement levels have a category $l \in \{1, 2, 3, 4\}$, a weighted Cohen's κ was performed that was 0.615 in the dataset. It shows a substantial agreement for recognising children's engagement levels. After data selection (same as in Chapter 4, see Section 4.2.5), there were 142416 frames and the number (percentage) of frames in each level of engagement were

shown in Table 5-5 from level 1 to level 4 respectively. OpenFace [9] was employed for extracting facial features in this study same as Chapter 4. For facial AU recognition, it is able to recognise a subset of AUs, specifically: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45 (see Figure 2-5). The intensity and occurrence of each facial AU would be used to measure the engagement levels of children during this test.

Level of Child Engagement	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
Count of frames (%)	9399 (6.60%)	14740 (10.35%)	87287 (61.29%)	30990 (21.76%)

Table 5-5. The distribution of frames in terms of child engagement levels, shown in the count of clips and in parenthesis in percent.

Classification

Children's facial action units were annotated by human labellers with four classes. Chapter 4 built a binary classification task to detect whether a child is engaged or disengaged in the digital stories viewing. Due to unbalanced class distribution and low accuracy of low engagement levels identification, that classifier was not used for the 4-class classification. As this study collects more data than in Chapter 4, a 4-class classification would be built as a multiple classification could reflect more information related to children's engaged states to the different media types of the story-stems. To perform the classification task, the LIBSVM library [19] was used as an efficient implementation of the standard soft-margin Support Vector Machine (SVM) [20]. In the first step, data were normalised by linear transformation into an [0,1] interval. Second, the SVM classifier was created to detect whether a child is engaged in the digital story-stem vignettes from the extracted facial action units. Like Chapter 4, the 70-30 ratio was chosen for frames with valid facial data in designing the classifier: 70% of the source data for training the model, and 30% of the source data for testing the model.

5.5 The Self-report Measure

Children were asked to fill in a questionnaire after the story completion in order to judge their level of engagement with the story and gather their opinions (i.e., likes and dislikes) about multimedia types. The questionnaire was taken from the Chapter 4 (see Table 4-8). Two items (Q7 and Q8) were added for assessing the aesthetic qualities related to story design using different media types. There were nine single choice questions (Q1- Q9) using a smiley-face based 5-point Likert scale [73]. Q10 and Q11 were two open-ended questions

at the end, which support children to describe their story experiences and attitudes. All questions of the questionnaire were shown in Table 5-6.

Questions	Aspects
Q1. I was absorbed in this story.	Distraction/attentional focus
Q2. I was involved in this story that I'm happy to tell people what happens next.	Distraction/attentional focus
Q3. I found this story confusing to understand.	Story understanding
Q4. When I was watching this story, I found myself thinking about other things.	Distraction/attentional focus
Q5. I was stressed while watching this story.	Empathy
Q6. I felt I knew what the child doll were going through emotionally in this story.	Empathy
Q7. I liked the dolls and images (doll house and furniture) used on this story.	Aesthetics
Q8. I liked the voice used on this story.	Aesthetics
Q9. I felt interested in this story task (including the story and this questionnaire).	General attitudes towards this task/Interest
Q10. How's the mummy doll feeling now? And what's the mummy doll thinking now?	Story understanding
Q11. How's the child doll feeling now? And what's the child doll thinking now?	Story understanding

Table 5-6. The items of the questionnaire and its related aspects.

For the five-point scale of this questionnaire, “not at all true” was coded as one and “very true” was coded as five. Children’s answers to each question were transformed to numerical scores (1~5) and analysed according to different media conditions (e.g. different storytelling voice, video format). The results will be presented in the following section.

5.6 Results

The results show the automatic recognition of child engagement levels under different multimedia types. Firstly, the performance of the classifier was evaluated that measure children’s engagement levels in the digital story-stem vignettes using the facial action units without considering different multimedia conditions. The following two subsections show the effects of children’s engagement levels on the two media types separately. Children’s engagement under the effects of four storytelling voice types were investigated firstly. Then the effect of children’s engagement levels on two presentation types (animations vs. live-action videos) were assessed. Lastly, children’s answers of the questionnaire were analysed to give a better understanding of their engaged states.

presence classifier, using a set of pairs (child engagement level/a set of AU presence) data. The accuracy of the AU intensity and AU presence classifier was also computed separately.

The accuracy of the AU intensity classifier was 66.14% (28257/42725 frames classified successfully) while the accuracy of the AU presence classifier was 60.94% (26037/42725 frames classified successfully). The accuracy of classifying child engagement into 4-class using the three facial AUs-related factors (i.e., AU intensity, AU presence and a combination of these two factors named as AU) shows that the highest accuracy rate of classification was the AU classifier, which means that the AU classifier had a better performance than the two other classifiers to measure children's engagement levels during following the digital story-stems on screen. A confusion matrix of the 4-class classification was firstly calculated as shown in Table 5-7 for the AU classifier, Table 5-8 for the AU intensity classifier and Table 5-9 for the AU presence classifier to display the distribution of child engagement levels. The AU classifier more successfully classified frames with not engaged (level= 1) and fully engaged (level= 4) categories than the other two classifiers. The AU intensity classifier correctly identified more frames with the highly engaged (level= 3) category than the other two classifiers while the AU presence had the best identification of frames with rarely engaged (level= 2) category.

Annotation Prediction	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
1 Not engaged	1852	1831	709	381
2 Rarely engaged	158	977	810	23
3 Highly engaged	644	4908	25388	2245
4 Fully engaged	102	862	1116	719

Table 5-7. Confusion matrix for the 4-class classification using the AU classifier, as shown in the number of frames.

Annotation Prediction	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
1 Not engaged	1634	1548	665	407
2 Rarely engaged	126	673	460	12
3 Highly engaged	891	5472	25605	2625
4 Fully engaged	107	873	1282	335

Table 5-8. Confusion matrix for the 4-class classification using the AU intensity classifier, as shown in the number of frames.

Annotation Prediction	1 Not engaged	2 Rarely engaged	3 Highly engaged	4 Fully engaged
1 Not engaged	1347	1315	591	25
2 Rarely engaged	359	1154	692	157
3 Highly engaged	904	5336	22878	2532
4 Fully engaged	156	768	3853	658

Table 5-9. Confusion matrix for the 4-class classification using the AU presence classifier, as shown in the number of frames.

To investigate the performance of the three AU-related classifiers, the 4-class classification task was transformed to multiple binary classification tasks. Four binary classifiers of engagement were constructed – one for each of the four engagement levels. The task of each of these classifiers is to discriminate a frame that belongs to engagement level l from a frame that belongs to some other engagement level $l' \neq l$, which was called $1 - v - other$, $2 - v - other$, etc.

Table 5-10 shows that the accuracy of the four binary classification given by the classifiers using facial AUs. All of the three architectures tested indicated that there was a difference in averaged performance metrics among the four tasks ($1-v-other$, $2-v-other$ etc.). The accuracy of each individual classifier shows that the four binary classification has a good accuracy performance for child engagement identification. However, the distribution of data was unbalanced, where the amount of “other levels l' ” class data was much greater than the engagement level l class. To prevent this problem of unbalanced class distribution, certain accuracy metrics used for the binary classification in Chapter 3 and 4 (see Table 3-9), such as such as precision, sensitivity/recall (true positive rate) as well as specificity (true negative rate), F₁ score, and Matthews correlation coefficient (MCC), were generalised to multi-class performance by averaging the performances of each individual class. While comparing the accuracy metrics, the AU classifier performed worse on the $2-v-other$ and $4-v-other$ than the other two classifiers. A low sensitivity in the two worse classifiers shows that the classifiers had a poor performance for identifying the frames with level 2 or level 4 correctly (0.0547 for the $2-v-other$ classifier and 0.1429 for the $4-v-other$ classifier). The confusion matrix (see Table 5-7) shows that a large number of frames were identified as highly engaged (level= 3) while labelling as rarely engaged (level= 2) from the $2-v-other$ classifier. The $4-v-other$ classifier had the similar result as the $2-v-other$ classifier.

Classifier	Acc.	BAcc.	Prec.	Spec.	Sens.	F ₁ score	MCC
1 – v – other	0.9183	0.7371	0.4000	0.9452	0.5290	0.4555	0.4170
2 – v – other	0.7882	0.5137	0.3357	0.9728	0.0547	0.0941	0.0619
3 – v – other	0.7924	0.7321	0.7925	0.5384	0.9258	0.8540	0.5209
4 – v – other	0.8776	0.5416	0.1696	0.9403	0.1429	0.1551	0.0900
<i>Avg.</i>	0.8441	0.6311	0.4245	0.8492	0.4131	0.3898	0.2725

Table 5-10. Accuracy metrics of the AU classifier for child engagement level $l \in \{1, 2, 3, 4\}$ using each of the three classification architectures. The avg. was the average value for the performances of each individual class.

Table 5-11 and Table 5-12 show that the four binary classification accuracy given by the classifiers using facial AU intensity and AU presence respectively. All of the three architectures tested indicated that there was a difference of averaged performance metrics across the four tasks (1-v-other, 2-v-other etc.). Like the AU classifier, the accuracy of two classifiers shows that the four binary classification has a good accuracy performance for engagement identification. While comparing the accuracy metrics, the AU intensity classifier had the similar performance with the AU classifier across the four classification tasks, where performed worse on the 2-v-other and 4-v-other than the other two classifiers. A low sensitivity in the two worse classifiers shows that the classifiers had a poor performance for identifying the frames with level 2 or level 4 correctly (0.0873 for the 2-v-other classifier and 0.0833 for the 4-v-other classifier). The confusion matrix (see Table 5-8) shows that a large number of frames were identified as highly engaged (level = 3) while labelling as rarely engaged (level = 2) from the 2-v-other classifier. The 4-v-other classifier had the similar result as the 2-v-other classifier.

Classifier	Acc.	BAcc.	Prec.	Spec.	Sens.	F ₁ score	MCC
1 – v – other	0.9179	0.6728	0.3711	0.9542	0.3913	0.3810	0.3371
2 – v – other	0.8034	0.5355	0.5725	0.9836	0.0873	0.1515	0.1648
3 – v – other	0.7835	0.7281	0.7932	0.5500	0.9061	0.8459	0.4999
4 – v – other	0.8678	0.5090	0.0982	0.9347	0.0833	0.0902	0.0195
<i>Avg.</i>	0.8432	0.6114	0.4588	0.8556	0.3670	0.3672	0.2553

Table 5-11. Accuracy metrics of the AU intensity classifier for child engagement level $l \in \{1, 2, 3, 4\}$ using each of the three classification architectures. The avg. was the average value for the performances of each individual class.

Due to similar performance between the AU classifier and the AU intensity classifier, the accuracy metrics were compared to find which classifier had a better performance. Firstly, the confusion matrix shows that the total number of frames that classified correctly across

the four engagement levels was higher using the AU classifier than the AU intensity one. The AU classifier had a higher number of frames that classified correctly in level 1, 2, and 4 and a slightly lower number in level 3 than the AU intensity classifier. Balanced accuracy, F₁ score, and MCC as three validation metrics for multi-class classification, show that the AU classifier had a better performance across the four levels (except in level 2) than the AU intensity classifier.

The AU presence classifier performed much better on the 3-v-other than the other classifiers, with both F₁ score and MCC. The sensitivity values indicated that a low proportion of frames were correctly identified across the four tasks. Only the 3-v-other had a better sensitivity value (0.7987) than other classifiers. The AU presence classifier has a much poor performance for low engagement levels (level 1 and 2) identification than the AU and AU intensity classifiers, as show in the validated metrics such as balanced accuracy and F₁ score. Thus, the AU presence classifier worked worse on the 4-class classification task than the other two classifiers.

Classifier	Acc.	BAcc.	Prec.	Spec.	Sens.	F₁ score	MCC
<i>1 – v – other</i>	0.9361	0.5054	1.0000	1.0000	0.0109	0.0215	0.1009
<i>2 – v – other</i>	0.7994	0.5142	0.5161	0.9912	0.0373	0.0695	0.0954
<i>3 – v – other</i>	0.7091	0.6686	0.7672	0.5384	0.7987	0.7827	0.3441
<i>4 – v – other</i>	0.8247	0.5360	0.1197	0.8786	0.1935	0.1479	0.0582
<i>Avg.</i>	0.8173	0.5561	0.6008	0.9566	0.2601	0.2554	0.1497

Table 5-12. Accuracy metrics of the AU presence classifier for child engagement level $l \in \{1, 2, 3, 4\}$ using each of the three classification architectures. The avg. was the average value for the performance of each individual class.

In a short summary, compared to the other two AU-related classifiers, the AU classifier had the best performance for the four binary classification task. For the 4-class classification, frames labelled as high engagement (the highly engaged and the fully engaged categories) contain less information related to fully engaged status as the AU classifier had a poor performance to distinguish the highly engaged and the fully engaged categories.

5.6.2 The Effect on Child Engagement of Storytelling Voices

The data used here were taken from the classification result of the AU classifier: a set of child-engagement-level/storytelling-voice-type pairs. Since both child engagement level and storytelling voice type are categorical variables, crosstabulation was computed along with a

Chi-Square²⁶ analysis. Crosstabulation is a statistical technique used to display a breakdown of the data by these two categorical variables. A Chi-Square test was performed to test where the results of a crosstabulation were statistically significant; that is, whether the two categorical variables (child engagement level and storytelling voice type) were independent or related to one another.

		Child Engagement Level				Row marginals
		Level 1	Level 2	Level 3	Level 4	
Storytelling Voice Type	Female Expressive	645	1988	7967	1193	11793
	Female Flat	665	2145	7020	773	10603
	Male Expressive	668	2196	7106	787	10757
	Male Flat	778	2249	5930	615	9572
Column marginals		2756	8578	28023	3368	42725

Table 5-13. Observed number of frames and marginals for the rows and columns by child engagement level and storytelling voice type. The marginals for the rows and columns were calculated by adding the frequencies across the rows and down the columns.

Table 5-13 shows counts of child engagement levels for the different storytelling voice types. For example, 665 frames taken from children's recordings of being asked to watch story-stems narrated by a female flat voice were classified as level 1. A Chi-Square test of independence was then performed to examine the relation between child engagement level and storytelling voice type. The null hypothesis (H_0) of the Chi-Square test was: child engagement level is independent of storytelling voice type.

In addition, a child engagement level * storytelling voice type crosstabulation was built (see Table 5-15) to check the *expected count* because the Chi-Square test cannot be used if the *expected count* was less than 5. The *expected count* value was the number of cases expected in each cell, calculated from the product of the row and column totals, divided by the total sample size. For example, the expected count in the top left cell (level 1, FE voice type) was 760.7, by calculating the product of row total (2756.0) and column total (11793.0) divided by total sample size (42725.0). Table 5-15 shows that no cells that have expected counts were less than 5 in this Chi-Square test, which means that this test is appropriate.

²⁶ <https://www.ncbi.nlm.nih.gov/books/NBK21907/>

The result was calculated using a Pearson Chi-Square statistic (χ^2), which involved the squared difference between the observed and the expected frequencies. Table 5-14 shows that there was a significant association between child engagement level and storytelling voice type ($\chi^2 (9, N=42725) = 314.961, p = .000$).

	Value	df	p-value (Asymptotic significance)
Pearson Chi-Square	314.961 ^a	9	.000
N of Valid Cases	42725		

Table 5-14. The result of Chi-Square Test. ^a: 0 cells had expected count less than 5. The minimum expected count is 617.45 (shown in the crosstabulation, Table 5-15).

Child Engagement Level * Storytelling Voice Type Crosstabulation

			Storytelling voice type				Row Totals
			FE	FF	ME	MF	
Child engagement level	1	Count	645 _a	665 _b	668 _b	778 _c	2756
		Expected Count	760.7	684.0	693.9	617.4	2756.0
		% with Engagement	23.4%	24.1%	24.2%	28.2%	100.0%
		% with Voice	5.5%	6.3%	6.2%	8.1%	6.5%
	2	Count	1988 _a	2145 _b	2196 _b	2249 _c	8578
		Expected Count	2367.7	2128.8	2159.7	1921.8	8578.0
		% with Engagement	23.2%	25.0%	25.6%	26.2%	100.0%
		% with Voice	16.9%	20.2%	20.4%	23.5%	20.1%
	3	Count	7967 _a	7020 _b	7106 _b	5930 _c	28023
		Expected Count	7734.9	6954.4	7055.4	6278.2	28023.0
		% with Engagement	28.4%	25.1%	25.4%	21.2%	100.0%
		% with Voice	67.6%	66.2%	66.1%	62.0%	65.6%
	4	Count	1193 _a	773 _b	787 _b	615 _c	3368
		Expected Count	929.6	835.8	848.0	754.6	3368.0
		% with Engagement	35.4%	23.0%	23.4%	18.3%	100.0%
		% with Voice	10.1%	7.3%	7.3%	6.4%	7.9%
Column Totals	Count	11793	10603	10757	9572	42725	
	Expected Count	11793.0	10603.0	10757.0	9572.0	42725.0	
	% with Engagement	27.6%	24.8%	25.2%	22.4%	100.0%	
	% with Voice	100.0%	100.0%	100.0%	100.0%	100.0%	

Table 5-15. The crosstabulation table between child engagement level and storytelling voice type (FE = female expressive, FF = female flat, ME = male expressive, MF = male flat). * Each subscript letter denotes a subset of storytelling voice type categories whose column proportions do not differ significantly from each other at the .05 level.

However, the Chi-Square cannot reveal how the two variables are related or how strong the relation is. A child engagement level * storytelling voice type crosstabulation was used (see Table 5-15) to display the frequency of different storytelling voice types broken down by child engagement level, with column percentage shows as the summary statistic. Besides the *expected count*, each cell includes: *count*, *row percentage (% with Engagement)*, and *column percentage (% with Voice)*. *Count* was the observed number of frames (same as values in

Table 5-13). *Row percentages (% with Engagement)* were expressed as a percentage of the level of engagement that each cell represents within a table row, calculated by dividing the cell count by the row total. *Column percentages (% with Voice)* were the percentage of the voice type that each cell represents within a table column, calculated by dividing the cell count by the column total. For example, in the top left cell (level 1, FE voice type), the row percentage was 23.4% (645 divided by 2756), which represents the percentage of frames from story-stems narrated by the FE voice and labelled as level 1 within all frames labelled as level 1. The column percentage was 5.5% (645 divided by 11793), which represents the percentage of frames labelled as level 1 from story-stems narrated by the FE voice within all frames narrated by FE voice type.

To investigate which storytelling voice type was a better type for engaging children in the story-stems, the engaged data was analysed (child engagement level 3: highly-engaged and 4: fully-engaged). For the row of engagement level 3, 28.4% of frames were from children's recordings when story-stems narrated by an FE voice. This compares to 25.1% of frames from story-stems narrated by an FF voice, 25.4% of frames of recordings using a ME voice and 21.2% of frames of recordings using a MF voice to present the story-stems. Similar to the row of engagement level 4, the highest percentage also occurred in the FE group (35.4%, which was 1193 divided by 3368). According to the *row percentage* of engaged data (both in level 3 and 4), a higher number/percentage of frames was for story-stems narrated by a FE voice than other voice types. This indicated that a female expressive storytelling voice was a better voice type for engaging children than other three voices.

The column proportions test assigns a subscript letter to the categories of the storytelling voice types. For each pair of columns, the column proportions are compared using a z test. If a pair of values is significantly different, the values have different subscript letters assigned to them. The percentages in the female flat storytelling voice type and male flat storytelling voice type categories both have the subscript 'b' as the percentages in those columns are not significantly different. However, the subscripts in the female expressive voice type (subscript 'a') and male flat voice type (subscript 'c') categories differ from each other as well as from the female flat and male expressive voice type (subscript 'b') categories. This means that the percentages in the female expressive storytelling voice type and male flat storytelling voice type categories are significantly different from each other as well as from the percentages in the female flat and male expressive storytelling voice type categories.

There was a significant association between child engagement levels and storytelling voice types. A female expressive storytelling voice was a better voice type for engaging children than other three voices. Although no significant differences were found between the FF and ME voice type, there was no need to consider distinguishing them because the aim of this chapter was to find the best voice type, and the performance of both were poorer than FE. Thus, the best storytelling voice type here was the female expressive voice type and the null hypothesis is rejected.

5.6.3 The Effect of Child Engagement on Presentation Types

The data used here were taken from the classification result of the AU classifier: a set of child-engagement-level/presentation-type pairs. Both child engagement level and presentation type are categorical variables. As for testing the relationship between child engagement level and presentation type, a crosstabulation with a Chi-Square test of independence was computed in this section. Table 5-16 shows counts of child engagement levels for the different presentation types. For example, 1432 frames taken from children's recordings of being asked to watch story-stems displayed with an animated video were classified as level 1.

		Child Engagement Level				Row marginals
		Level 1	Level 2	Level 3	Level 4	
Presentation Type	Animation	1432	4483	15622	1956	23493
	Live-action	1324	4095	12401	1412	19232
Column marginals		2756	8578	28023	3368	42725

Table 5-16. Observed number of frames and marginals for the rows and columns by child engagement level and presentation type. The marginals for the rows and columns were calculated by adding the frequencies across the rows and down the columns.

A Chi-Square test of independence was then performed to examine the relationship between child engagement level and presentation type. The null hypothesis (H_0) of the Chi-Square test was: child engagement level is independent of presentation type. Since the Chi-Square test cannot be used if the *expected count* was less than 5, a child engagement level * presentation crosstabulation was built (see Table 5-18) to check the *expected count* of each cell. The *expected count* value in each cell was the product of the row and column totals, divided by the total sample size, if the variables were statistically independent. Table 5-18 shows that no cells that have expected counts were less than 5 in this Chi-Square test, which means that this test is appropriate. The results of the Chi-Square test (see Table 5-17) shows

that there was a significant association between child engagement level and presentation type (χ^2 (3, N = 42725) = 55.474, $p = .000$).

	Value	df	p-value (Asymptotic significance)
Pearson Chi-Square	55.474 ^a	3	.000
N of Valid Cases	42725		

Table 5-17. The result of Chi-Square Test. ^a: 0 cells (0.0%) have expected count less than 5. The minimum expected count is 1240.6 (shown in the crosstabulation, Table 5-18)

However, the Chi-Square value cannot reveal how the two variables are related or how strong the relation is. A child engagement level * presentation type crosstabulation was built (see Table 5-18) to display the frequency of different presentation types broken down by child engagement level, with column percentage shows as the summary statistic. Across presentation types, the column totals show that 55.0% of frames of children's recordings was from animated story-stems while 45.0% of frames of recordings was for story-stems displaying with live-action-recorded videos.

Child Engagement Level * Presentation Type Crosstabulation

			Presentation type		Row Totals
			Animation	Live-action	
Child engagement level	1	Count	1432 _a	1324 _b	2756
		Expected Count	1515.4	1240.6	2756.0
		% with Engagement	52.0%	48.0%	100.0%
		% with Presentation	6.1%	6.9%	6.5%
	2	Count	4483 _a	4095 _b	8578
		Expected Count	4716.7	3861.3	8578.0
		% with Engagement	52.3%	47.7%	100.0%
		% with Presentation	19.1%	21.3%	20.1%
	3	Count	15622 _a	12401 _b	28023
		Expected Count	15408.9	12614.1	28023.0
		% with Engagement	55.7%	44.3%	100.0%
		% with Presentation	66.5%	64.5%	65.6%
	4	Count	1956 _a	1412 _b	3368
		Expected Count	1851.9	1516.1	3368.0
		% with Engagement	58.1%	41.9%	100.0%
		% with Presentation	8.3%	7.3%	7.9%
Column Totals		Count	23493	19232	42725
		Expected Count	23493.0	19232.0	42725
		% with Engagement	55.0%	45.0%	100.0%
		% with Presentation	100.0%	100.0%	100.0%

Table 5-18. The crosstabulation table between child engagement level and presentation type. * Each subscript letter demotes a subset of presentation type categories whose column proportions do not differ significantly from each other at the .05 level.

To investigate which presentation type was a better type for engaging children in the story-stems, the engaged data was analysed (child engagement level 3: highly-engaged and 4: fully-engaged). The row of child engagement level 3 shows that 55.7% of frames were from children's recordings when they were watching the animated story-stems. This compares to 44.3% of frames of recordings from story-stems displayed as live-action videos. A higher percentage also occurred in the animation type from the row of engagement level 4 (58.1% for animation and 41.9% for live-action videos). According to the *row percentage* of engaged data (both in level 3 and 4), the number of frames from the animation presentation type was higher than from the live-action type. Combined with the result of Chi-square test, this indicated that animated story-stems was a better presentation style than the live-action videos.

The column proportions test assigns a subscript letter to the categories of the two presentation types and the two column proportions are compared using a z test. If a pair of values is significantly different, the values have different subscript letters assigned to them. The subscripts in the live-action presentation type (subscript 'a') and animation presentation type (subscript 'b') categories differ from each other. This means that the percentages in the live-action presentation type and animated presentation type categories are significantly different from each other.

Thus, there was a significant association between child engagement levels and presentation types. The better presentation type between the two types was the animated presentation type and the null hypothesis is rejected.

5.6.4 The Self-report Measure

Each child was asked to fill in a questionnaire (see Appendix F) after completing each MCAST story vignette. There are 4 MCAST story-stems so that each child was asked to complete 4 questionnaires during the whole test. Children's answers were used to collect their opinions for the use of different multimedia types. Like the questionnaire of Chapter 4, there were two open-ended questions (Q10 and Q11) about the feelings of the child doll and the mummy doll respectively in relation to story understanding. For other questions, nine single choice questions (Q1- Q9, see Table 5-6) used a smiley-face based 5-point Likert scale to investigate the five aspects of child engagement. "Not at all true" using a totally sad face was coded as one and "really true" using a totally happy face was coded as five.

There were two steps for analysing children's answers of this questionnaire. Firstly, an overall descriptive statistical analysis of children's answers was performed according to the five aspects of children engagement measurements. For the second step, the children's answers were analysed according to the two key media conditions studied in this chapter: storytelling voice type and presentation type.

Aspects	Related Questions	Number of children	Mean	S.D.
Distraction/attentional focus	Q1, Q2, Q4	40	3.63	1.198
Story understanding	Q3	40	4.17	0.892
Empathy	Q5, Q6	40	2.93	1.385
Aesthetics	Q7	40	3.40	1.264
	Q8	40	3.66	1.177
General attitudes/Interest	Q9	40	3.75	1.346

Table 5-19. Descriptive analysis of children's answers for the questionnaire according to the five aspects of child engagement measurements.

Error! Reference source not found. shows the descriptive analysis of children's answers according to the five aspects of child engagement measurements. Firstly, the aspect of distraction/ attentional focus aims to measure the extent of children's concentration and absorption in each story-stem. The results of children's answers to questions related to this aspects (3.63/5) show that children were able to pay attention to the story-stems and sometimes were distracted by other things. Then for the aspect of story understanding, a sad face means that the story was really confusing to understand while a totally happy face means that the story was very easy to understand. Children's answers show that all the four MCAST story-stems were easy to understand and 35% of children's answers were coded as 5, which means that stories were quite easy to understand for them. The third aspect in the questionnaire was empathy, which aims to measure if a child participant can feel with the child doll's distressed emotions due to a predicament shown in each story-stem. The average score of children's answers of Q5 was 2.74/5 (S.D.=1.531), which means that children did not think they felt distressed like the child doll while watching the MCAST story-stems. Their answers of Q6 shows that children think they can feel with the child doll's emotion represented as distress (3.13/5, S.D. = 1.196). The next aspect was called aesthetics, which focuses on children's like or dislike for the multimedia elements including animation, live-action video and storytelling voice. Overall, their answers of this aspect indicate that children have a high level of emotional engagement during following the redesigned stories on screen. A detailed analysis will be performed in the next step. From the last aspect, general

attitude/interest, the whole test (including story viewing, story completion and filling in a questionnaire) was interesting for child participants (3.75/5 for Q9). 38% of children chose “Yes, I really like them!” (coded as 5) for this test. For children who dislike the test (scores under 3), 12.5% of children chose “No, I don’t like them at all!” (coded as 1) and only 5% chose “No, I don’t like them.” (coded as 2). This indicated that children had a strong attitude to express their dislike.

Then the second step focuses on analysing children’s answers according to the two key media conditions studied in this chapter: storytelling voice types and presentation types. The aspect for analysing the two media conditions in the questionnaire was aesthetics. Two questions (Q7 and Q8) in the aspect of aesthetics: Q7 was the item to ask the presentation types while Q8 was the item to ask the storytelling voice types.

The analysis of children’s answers under different storytelling voice types

The first part in this step was to analyse children’s answers according to the four storytelling voice types. A descriptive analysis for the five aspects of child engagement measurements according to different storytelling voice types was shown in Table 5-20. In the aspect of Aesthetics, Q8 was only analysed as it focuses on storytelling voices. Since child participants were allocated to watch the MCAST stories displayed in different media types, children’s answers were grouped according to the storytelling voice types. For example, to analyse children’s answers for the ‘Female Flat’ voice type, the answers of the questionnaire were collected from the following child participant groups for the four MCAST story-stems: the ‘Nightmare’ story-stem of Group 1 and 5, the ‘Hurt Knee’ story-stem of Group 3 and 7, the ‘Illness’ story-stem of Group 2 and 6, and the ‘Shopping’ story-stems of Group 4 and 8 (Full details see Table 5-4). Since each story-stem contains 2 groups (10 child participants), 40 questionnaires were collected for each storytelling voice type.

Aspects	Female Flat	Female Expressive	Male Flat	Male Expressive
Distraction/attentional focus	3.64(1.187)	3.67(1.183)	3.58(1.193)	3.63(1.243)
Story understanding	4.23(0.802)	4.27(0.751)	3.90(1.150)	4.27(0.784)
Empathy	2.98(1.418)	3.00(1.414)	2.76(1.265)	2.97(1.445)
Aesthetics (Storytelling Voice)	3.60(1.057)	3.90(1.008)	3.53(1.320)	3.61(1.297)
General attitudes/Interest	3.75(1.335)	3.88(1.324)	3.55(1.431)	3.83(1.318)

Table 5-20. Descriptive analysis for children’s answers for each aspect under the different storytelling voice types, shown with mean values and in parenthesis in std. Dev.

Besides the analysis of mean and variance values, a Kruskal Wallis²⁷ test was then used to test for children's answers to each question respectively among the four storytelling voice types. The reason of using the Kruskal Wallis test was the nature of variables²⁸: children's answers (dependent variables) are measured as the ordinal level and the storytelling voices (independent variables) consist of 4 categorical, independent groups. The null hypothesis (H_0) of each question was: there is no significant association of children's answers to a specific question among different storytelling voice types.

The aspect of distraction/attentional focus includes Q1, Q2 and Q4. Children's answers to these three questions was firstly calculated by mean value and variance. The results show that there was not much difference in mean and variance for children's answers among different storytelling voice types, went from 3.67 to 3.58 with a standard deviation at ~1.1. The three questions were then analysed respectively. Q1 asks children's absorption in the story-stem and children's average rating of Q1 was 3.62/5 (S.D.=1.148) in the female expressive voice condition, a higher rating result than other storytelling voice types (e.g., 3.35/5 (S.D.=1.231) in the male flat voice type). The result of the Kruskal Wallis test shows there were no significant differences in children's answers of Q1 according to the storytelling voice types ($H(3) = 1.336$, $p = .721$) and the null hypothesis for Q1 is accepted. Q2 focuses on whether children were happy to complete the MCAST story-stem. The average rating of Q2 shows that the highest average one from the female expressive voice (4.10, S.D.=1.127). The result of the Kruskal Wallis test shows there were no significant differences in children's answers of Q2 according to the storytelling voice types ($H(3) = 0.958$, $p = .811$) and the null hypothesis for Q2 is accepted. The completed story and children's behaviours will be used for attachment assessment. If a child completes the story spontaneously, MCAST assessors or the SAM system can collect more reliable data to evaluate the child's attachment status. However, children's answers indicated that their intention to spontaneously complete the story was not changed according to different storytelling voice types. Similar to Q1 and Q2, the result of the Kruskal Wallis test of Q4 shows there were no significant differences in children's answers according to the storytelling voice types ($H(3) = 0.576$, $p = .902$). Thus, the results of the Kruskal Wallis test

²⁷ <https://stats.idre.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#kw>

²⁸ <https://stats.idre.ucla.edu/other/mult-pkg/whatstat/>

indicated that children do not think their attention/distraction would have a significant difference according to different storytelling voice types.

The aspect of story understanding only includes Q3. In Q3, a sad face displayed on the questionnaire means that the story was really confusing to understand while a totally happy face means that the story was very easy to understand. The average rating results of Q3 show that the story-stems narrated by a MF voice were more difficult to understand for children than those narrated by other voices (3.9 in a MF voice while ~4.2 in other voice types). Under the expressive voice conditions, the average ratings of children's answers of Q3 were the same between the different voice genders (4.27/5). However, the result of the Kruskal Wallis test shows the null hypothesis for Q3 is accepted as children do not think their comprehension was associated with the storytelling voice types ($H(3) = 2.852, p = .415$).

The third aspect in the questionnaire was empathy, which aims to measure if a child participant can feel with the child doll's emotion represented as distress due to a predicament shown in each story-stem. The average rating results of children's answers of Q5 under all the four storytelling voice types were less than 3, with 5 being the top possible rating. Both Kruskal Wallis tests for Q5 and Q6 show there were no significant differences in children's answers according to the storytelling voice types (Q5: $H(3) = 0.220, p = .974$, Q6: $H(3) = 4.954, p = .175$) and the null hypotheses for these two questions are accepted. This means that children did not think they can feel with the child doll's distressed emotions when watching the MCAST story-stems narrated by any of the four storytelling voice types and they also did not think their empathy could be improved by using different storytelling voice types.

The next aspect in this part, aesthetics, aims to analyse if children like or dislike the storytelling voice they listened (Q8). The average rating of Q8 was 3.9/5 (S.D.=1.008) in the female expressive voice types, slightly higher than the ratings (~3.6) in the other three voice types. However, the result of Kruskal Wallis tests show there were no significant differences in children's answers of Q8 and storytelling voice type ($H(3) = 1.917, p = .590$) and the null hypothesis is accepted. This means that children do not have their favourite storytelling voice type and all four storytelling voice were acceptable for them (The mean values were larger than 3 among all four storytelling voice.)

The last aspect aims to measure children's general attitude/interest towards the whole test (including story viewing, story completion and filling in a questionnaire). The result of Kruskal Wallis tests show there was not a significant association between children's answers of Q9 and storytelling voice type ($H(3) = 1.343$, $p = .719$) and the null hypothesis of Q9 is accepted. This indicates that children's ratings towards their attitudes/ interest would not be related to the four storytelling voice types.

In a short summary, the comparison of the mean and variance values indicates that the female expressive voice had a better performance for engaging children. However, the results of Kruskal Wallis tests show that there was not a significant association between children's answers to all questions and storytelling voice types and the null hypotheses of all questions were accepted. Therefore, children's answers indicated that they do not think their engagement levels would be affected when they were watching the story-stems narrated by different storytelling voices.

The analysis of children's answers under different presentation types

The second part in this step was to analyse children's answers according to the two presentation types. Since child participants were allocated to watch the MCAST stories displayed in different media types (see Table 5-4), children's answers were grouped according to the presentation types: Group 1-4 (20 children) was asked to watch the animated MCAST videos while Group 5-8 (20 children) for the live-action recorded MCAST videos. A descriptive analysis for the five aspects of child engagement measurements according to different presentation types is shown in Table 5-21. In the aspects of Aesthetics, Q8 was only analysed as it focuses on presentation types.

Aspects	Animated Video	Live-action Recorded Video
Distraction/attentional focus	3.65(1.211)	3.61(1.188)
Story understanding	4.27(0.7627)	4.06(0.998)
Empathy	2.99(1.412)	2.88(1.359)
Aesthetics (Video Format)	3.48(1.263)	3.30(1.267)
General attitudes/Interest	3.78(1.319)	3.72(1.380)

Table 5-21. Descriptive analysis for children's answers for each aspect under the different video formats, shown with the mean values and in parenthesis in std. Dev.

Besides the analysis of mean and variance values, a Kruskal Wallis test was then used to test for children's answers to each question respectively among the two presentation types. The

reason of using the Kruskal Wallis test was the nature of variables: children's answers (dependent variables) are measured as the ordinal level and the video formats (independent variables) consist of 2 categorical, independent groups. The null hypothesis (H_0) of each question was: there is no significant association of children's answers to a specific question among different storytelling voice types.

The aspect of distraction/attentional focus includes Q1, Q2 and Q4. Children's answers between the animated and the live-action recorded video format show that the average ratings were similar (difference only .04 with std. Dev of ~1.1) between the two video formats. The three questions were then analysed respectively. Q1 was related to children's absorption and children's average rating was 3.51 (S.D.=1.191) in the live-action presentation type, slightly higher than the average rating in the animation type (3.40/5, S.D.=1.228). However, the average ratings of children's answers of Q2 and Q4 was not consistent with answers of Q1. The average rating results of Q2 and Q4 show that the animation type had a higher average rating of children's answers than the live-action type. The statistical results of these three questions show that there were no significant differences in children's answers according to the presentation types under a Kruskal Wallis test (Q1: $H(1) = 0.371$, $p = .542$, Q2: $H(1) = 0.708$, $p = .400$, and Q4: $H(1) = 0.431$, $p = .511$) and all three null hypotheses are accepted. This means that children's attentional focus/ distraction was not affected by video formats. For example, children do not think an animated video can hold their attention for longer.

Children's answers to Q3 show the result of the aspect of story understanding. The average rating results of Q3 show that the story-stems displayed with animated videos (4.27/5) were more easily to understand for children than displayed with live-action videos (4.06/5). The result of the Kruskal Wallis test shows the null hypothesis for Q3 is accepted as children do not think their comprehension was associated with the presentation types ($H(1) = 2.852$, $p = .415$).

The third aspect in the questionnaire was empathy including Q5 and Q6. The average ratings of Q5 show that children did not think they can feel with the child doll's distressed emotions in a predicament shown in each story-stem displayed with either of video format, with an average rating of 2.72/5 for the live-action video, and 2.76/5 for the animated video. Q6 was analysed whether children could know what the child doll were going through emotionally in this story. The average rating of Q6 shows that children cannot understand the child doll's

emotion represented as distress. Meanwhile, the result of two Kruskal Wallis tests for Q5 and Q6 show there were no significant differences in children's answers according to the presentation types (Q5: $H(1) = 0.001$, $p = .982$, Q6: $H(1) = 1.094$, $p = .296$) and the null hypotheses for these two questions are accepted. This means that children did not think they can feel with the child doll's distressed emotions when watching the MCAST story-stems and their also did not think their empathy could be improved by watching the story-stems displayed with different presentation types.

The aspect of aesthetics in this part focuses on analysing whether children like or dislike the videos they watched (Q7). The overall average rating of children's answers was 3.393 (S.D.=1.264), with 5 being the top possible score. Compared to the two presentation types, there was not much difference in children's answers, with an average rating of 3.48/5 for the animated types and 3.3/5 for the live-action type. The result of Kruskal Wallis tests show there were no significant differences in children's answers according to the presentation types (Q7: $H(1) = 0.957$, $p = .328$) and the null hypothesis is accepted. This means that children cannot say which presentation type is more attractive to them and both animation and live-action were acceptable for them (Both mean values were larger than 3.)

The last aspect aims to measure children's attitudes/interest towards the whole test (including the story viewing and questionnaire). The average rating of Q9 indicates that children felt interested in watching the story-stems displayed with both presentation types. Meanwhile, there were no significant differences in children's answers of Q9 according to the presentation types under a Kruskal Wallis tests ($H(1) = 0.050$, $p = .824$) and the null hypothesis of Q9 is accepted. This indicates that children's ratings towards their attitudes/interest would not be related to the video formats for displaying the story-stems.

In a short summary, the comparison of the mean and variance values and the results of Kruskal Wallis tests indicate that there was not a significant association between children's answers to all questions and video formats. Also, the null hypotheses of all questions were accepted. Therefore, children do not think any aspect related to their engagement would be affected when watching the story-stems displayed with animation or live-action videos.

5.7 Discussion

The digital story-stem approach has been recognised as a reliable and cost-effective method for child psychiatric studies. But providing such tests via computer relies on the child being

engaged in the story. An engaging story could make children improve their attentional focus towards the story and its comprehension. Digital stories can be constructed using a mixture of graphics, animation, text, recorded audio narration, video and music, to present information on a specific topic. Two important media types, including a voice type and a presentation type, of digital were studied in this chapter. This chapter was focused on investigating two key aspects: storytelling voice and the video format, to create an engaging MCAST digital story-stem vignettes to help children get absorbed to complete the vignette in spontaneous play, to answer RQ2: *How do voice type and presentation type affect child engagement levels in digital story-stems?*

Two conditions of the storytelling voice were: voice gender (female vs male) and voice expressiveness (expressive voice vs flat voice). For the video format, MCAST stories were redesigned using animation tools narrated by the above storytelling voices. Animated MCAST stories were used to compare it to the live-action MCAST video used in Chapter 4.

Children's engagement levels could reflect the extent of how attentive and attractive of the MCAST stories with different media types for them. The methods for measuring children's engagement were taken from Chapter 4. Results from Chapter 4 show that children's engagement levels (High vs. Low Engagement) could be measured using their spontaneous facial expressions, which was used to answer RQ1. However, a binary classification was not enough to provide information related to children's engagement, such as the percentage of fully engagement time during the engaged period. Therefore, this chapter builds a 4-class classifier to classify children's engagement levels. Facial data from 40 children were collected using an RGB webcam while they watched the story-stems from MCAST. The procedure of data collection and selection was the same as in Chapter 4.

There were three facial AU-related classifiers using the AU intensity, AU presence and a combination of these two factors named as AU respectively. The accuracy value of the three AU-related classifiers respectively shows a good performance of the 4-class classification, correctly in over 60% of instances and the highest accuracy rate of classification from the AU classifier. Confusion matrices were then computed for the three AU-related classifiers to display the distribution of child engagement. For example, the confusion matrix calculated from classification results using the AU classifier shows that the percentage of frames that correctly classified into all testing frames across the four engagement levels was 4.33% (1852), 2.29% (977), 59.42% (25388), 1.68% (719) from the level 1 to 4 respectively. It

indicates that the distribution of classified data was unbalanced. In addition, there was a number of frames classified as level 3 (highly-engaged) while labelling as level 2 (rarely engaged, 4908 frames) as well as level 4 (fully engaged, 2245 frames) using the AU classifier. It means that 11.5% of instances (4908/42725 frames) was wrongly classified between 'high engagement' and 'low engagement' and 5% of instances was difficult to distinguish the extent of high engagement, between level 3 and 4. The AU-intensity and AU-presence classifiers had a lower accuracy than the AU classifier. Compared to the confusion metrics calculated from the AU-intensity and AU-presence classification, the current results show that the AU classifier, the best one of AU-related classifiers, had an overall good performance of multi-class classification but relatively poor identification on level 2.

Due to unbalanced distribution of the data, the accuracy metrics were then computed to report the classification performance. Both the AU classifier and the AU intensity classifier had similar performance and better than the performance of the AU presence classifier. The accuracy metrics were compared to find which classifier had a better performance between the AU classifier and the AU intensity classifier. The AU classifier correctly identified more frames than the AU intensity classifier from the accuracy. Besides the accuracy, the values of balanced accuracy, F_1 score, and MCC as three validation metrics for multi-class classification, also show that the AU classifier had a better performance across the four levels (except in level 2) than the AU intensity classifier.

Chapter 4 found that facial AUs contained information related to children's engagement levels and demonstrated by building a binary classifier which identified child engagement between high and low engagement correctly in about 66% of cases with a balanced accuracy value. Compared to results in Chapter 4, the accuracy metrics in this chapter indicated that facial AUs contained information correlated with the four levels of child engagement. The 4-class classification task had a lower accuracy than the binary classification in Chapter 4, correctly in 38.98% of instances for the 4-class classification and of 87.99% for the 2-class classification with an F_1 score for both. Therefore, a binary classification using the facial AU intensity, which were taken from large amounts of video of spontaneous actions, can be a more reliable method for measuring children's engagement in 2 levels (high vs low) than the 4-class engagement identification in the context of digital story-stems.

The second step was to investigate the effect of multimedia types in digital story-stems on children's engagement levels. The first part focuses on investigating the effect of the

storytelling voice on child engagement levels. Although only one female and male storyteller recruited to display the MCAST stories in this study, both storytellers have rich experience in storytelling and also attended a brief MCAST administration training. Thus, they could provide a good quality of narration for the MCAST story-stems. The pitch value was used to distinguish the four storytelling voices and ensure that the expressive recordings were sufficiently more emotional and expressive than the flat recordings. Although a one-way ANOVA table shows that the four storytelling voices had a significant difference across the four child engagement levels, the results of the *post hoc* test indicated that there was a significant difference between female flat and expressive voice types ($p < .001$), but not a significant difference between the male flat and male expressive voice types ($p = 0.308$). When the male storyteller narrated the story-stem in his expressive voice, the pitch value was similar to a female storyteller's flat voice. This indicates that the male storyteller increases his pitch to express the voice expressiveness when narrating the story-stems.

The distribution of child engagement according to different storytelling voice types was compared to find the best storytelling voice type for story design. The crosstabulation along with a Chi-Square test shows that there was a significant association between child engagement levels and storytelling voice types ($p = .000$) and the null hypothesis is rejected. To analyse the engaged data (child engagement level in 3 and 4), the *row percentage (% with Engagement)* from crosstabulation shows that more engaged data was collected when children were watching story-stems narrated by a female expressive storytelling voice. Meanwhile, the column proportions test by assigning a subscript letter to each voice type indicated that the percentages in the female expressive storytelling voice type are different from the other three voice types. Therefore, a female expressive storytelling voice was a better voice type for engaging children than other three voices.

The second part focuses on investigating the effect of presentation types, animation vs live-action recorded video, on child engagement levels. The crosstabulation along with a Chi-Square test shows that there was a significant association between child engagement levels and presentation types ($p = .000$). The percentages of engaged data (both in level 3 and 4) show that a higher percentage occurred in the animation type than the live-action video type (74.8% vs 71.8%). Also, the subscripts to the categories of the animation and the live-action presentation type are different, which means that the percentages in the two presentation types differ from each other. Thus, animation can engage children better than the live-action videos. This study suggested that animation was a good alternative presentation type for the

design of the digital story-stems for the MCAST test so that to bring children into a deep engagement with the digital MCAST story-stems.

The overall rating results from the questionnaire show that children think they were able to understand and pay attention to story-stems while they did not think they can feel with the child doll's distressed emotion represented in the story-stems well during watching the digital story-stems. These findings were similar to results of the same questions in Chapter 4. As this study was not focused on the analysis of children's frequent facial actions, the relationship between their answers and the analysis of facial action units was unknown.

Children's answers of questionnaires were divided into two parts to investigate two media conditions respectively. According to different storytelling voice types, the analysis of the average ratings of children's answers to questions indicated that children had a slightly better performance from the story-stems narrated by a female expressive voice than narrated by other storytelling voice types. However, the comparison of the mean and variance values of children's answers between the two presentation types indicated that children do not think their engagement would be affected from the story-stems displayed with animation or live-action videos. Moreover, a Kruskal Wallis test was then used to detect if there was a significant association between children's answers to each question respectively and the storytelling voices/ video formats. The results of the Kruskal Wallis test indicated that there were no significant differences in children's answers to each question according to the storytelling voices as well as video formats respectively.

The result of questionnaire was not the same as results from the analysis of facial action units as children do not think their engagement could be improved when watching the story-stems created by the media types (storytelling voice and video format) used here. This may be because children's answers to questionnaire relies on their interpretation of questions. For studies involving children, Hanna *et al.* suggested considering observed facial expressions of children, such as frowns and yawns, as a better engagement indicator rather than their answers to questionnaires [43]. Therefore, the spontaneously facial expressions of children can be recognised as more reliable data to be used for measuring children's engagement levels in this thesis.

5.8 Conclusion

This chapter focuses on investigating the effect of two multimedia aspects, the storytelling voice and video format, on children's engagement levels to create an engaging digital MCAST story. The main contribution of this chapter was to find the best way of creating an engaging story was a combination of animation and a female expressive voice, which was answered RQ2.

The engaging digital story-stems makes children more easily get absorbed in the MCAST test. The best way of creating an engaging digital story could be used to design other story-stems for child psychiatric studies. Children were more engaged with a female expressive storytelling voice because a female expressive voice can better express both a mother's emotion and the predicament in the MCAST story-stems to improve children's attention and help them locate themselves as the "child doll" within the story-stem. While comparing to the live-action recorded MCAST videos in Chapter 4, the animated video was a more engaging presentation type as it makes children more easily get absorbed in the story-stems.

Meanwhile, the methods for measuring the child engagement levels were taken from Chapter 4. Since a larger dataset was collected in this chapter than in Chapter 4, a further step of identifying child engagement levels with a 4-class classification was conducted. The accuracy metrics show that the performance of the 4-class classification task for child engagement levels was poorer than the performance of the binary classification in Chapter 4. Thus, a binary classification would be more recommended as it is more reliable than a 4-class classification and it is able to detect whether a child is engaged in the digital story-stem vignettes from the extracted facial action units. This gives a support to answer the RQ1.

Chapter 6 Discussion and Conclusions

6.1 Introduction

This thesis focuses on measuring children's engagement levels in digital story-stems and investigating the effects of multimedia tools on creating a better and more engaging digital story. This chapter discusses three aspects corresponding to the two research questions: 1) measuring children's engagement levels using their facial behaviours; 2) designing an engaging story for children.

6.2 RQ1: Can children's spontaneous facial expressions be used to automatically measure engagement levels in digital story-stems?

The problem of automatic engagement measurement in a specific task is an area of great interest across a wide variety of fields. Children's engagement has been growing recognition of the importance of educational systems because monitoring children's engagement states is beneficial for adjusting the learning process. However, apart from education, little is known about children's engagement in other contexts.

The important area related to engagement in this thesis is to measure children's engagement levels in digital story-stems. The story-stem approach is a reliable and valid method for investigating the important relationships in a child's life and contributing to the Attachment theory. Engagement is an important concept in Attachment tests using story-stems; bringing the child into a deep engagement with a story is a key step to bring out his/her mental representation of attachment. The instance of the story-stem approach used in this thesis is the *Manchester Child Attachment Story Task* (MCAST) [32]. In the MCAST test, engagement is measured by a trained assessor's observation of facial expressions, using the MCAST protocol.

To reduce the cost and human involvement of child engagement assessment, one kind of automatic engagement recognition, based on computer vision, provides an automatic

identification of engagement by analysing cues from the face and gestures. These kinds of behaviours can be collected in a non-invasive way with simple sensors. Chapter 4 used two methods, gaze behaviours and facial action units, to successfully measure child engagement levels in the context of digital MCAST story viewing, which was answered RQ1: *Can children's spontaneous facial expressions be used to automatically measure engagement levels in digital story-stems?*

From the literature, eye-tracking has been used to measure engagement in previous studies. However, it was mainly trained on adults and there is little research related to the analysis of children's eye behaviours. To begin the research, an initial study in Chapter 3 was undertaken on adult engagement in digital story-stems to test basic eye-tracking technique, which would then be used with children in Chapter 4. Chapter 3 found that there were significant differences in fixation duration across distinct levels of engagement in digital story-stems and fixation duration is a good indicator that can be used to classify adult engagement. Meanwhile, Chapter 3 also developed an annotation scale based on adults' engagement behaviours. The annotation scale was used for coding children's engagement levels in Chapter 4. In Chapter 4, the procedure for recognition of child engagement using the eye-tracking technique showed that fixation was the primary eye-tracking feature of child engagement in digital story-stems. The descriptive statistics of fixation metrics showed that both the average number of fixations per clip and the mean fixation durations increased according to the increased levels of child engagement. However, the total number of fixations and total fixation duration increased when the engagement level increased from level 1 to 3 while decreased when the engagement level increased from level 3 to 4. Moreover, the results of the ANOVA and its *post hoc* test show that there were statistically significant differences in fixation duration across distinct levels of child engagement in digital story-stems so that fixation duration is a good indicator that can be also used to classify child engagement.

The eye-mind assumption in reading indicated that the eye remains fixated on a word as long as the word is being processed [90]. Thus, a corollary to this assumption, used in this thesis, was that the length of fixation durations on screen can reflect the extent of understanding during following of the story-stem video on screen. Longer fixation durations in a clip can reflect that a child was deeply understanding the story-stems. A deeper understanding is a good indicator for narrative engagement as introduced in the literature, where a child could locate him or herself within the mental model of the story-stem.

Since the results of fixation metrics indicated fixation duration contained information related to children's engagement levels, a binary classification task was performed where each clip of children's recordings is marked as being either high or low engagement levels using fixation duration. The classifier measured engagement correctly in 78% of instances with an F_1 score, a good result for automatic child engagement identification.

Besides eye-tracking, another method tested for child engagement measurements was coding children's facial expressions in terms of facial action units, such as brow movements, nose wrinkle, chin raise, and lip actions. Facial AUs have been used to measure children's engagement in the context of problem-solving [50]. Based on human annotation and selection results in terms of child engagement, the high versus low level of child engagement was successfully recognised as a binary classification using a set of child-engagement/facial-action-units pairs, in which not and rarely engaged levels were grouped into a 'low engagement' class, and highly and fully engaged categories were grouped into a 'high engagement' class.

The performance of the classification tasks was calculated by accuracy metrics and the results of accuracy metrics show that the best classifier between the three facial AU-related classifiers (i.e., AU intensity, AU presence and a combination of these two factors named as AU) was the AU intensity classifier. It had a better performance than the two other classifiers to identify children's engagement (correctly in about 87% of cases with an F_1 score) in the digital MCAST story-stems. The frequency of facial Action Units was then analysed using the frames identified with high engagement using the AU intensity classifier. The most frequent facial action units included facial movements of eyebrow (AU01, 02), mouth (AU12, 14) and chin (AU17). Frames with high intensity of AU02 and 14 are typically classified into high engagement levels using the AU intensity classifier. The eyebrow raise (AU02) and the horizontal mouth stretch movement (AU14) are two good indicators for fearful facial expressions [24, 35]. Based on the context of the MCAST test, the frequent AU02 and AU14 could explain that a child was engaging with the situation of specific anxiety and distress in the MCAST story, because feelings of anxiety facially corresponded to the elements of the expression of fear [35].

Moreover, as this thesis developed a scale for engagement annotation with 4 different levels, a 4-class classification was also built in Chapter 5 to test how much the facial action units contain information related to children's engagement levels. The accuracy metrics show that

the performance of the 4-class classification for child engagement levels was poorer than the performance of the binary classification, correctly in 38.98% of instances for the 4-class one and 87.99% of instances for the binary one with an F_1 score for both. Although a 4-class classification task could identify the extent of children's engagement, a binary classification would be more recommended as it is more reliable than a 4-class classification and it is able to detect whether a child is engaged in the digital story-stem vignettes from the extracted facial action units.

Since fixations (pauses over informative regions of interest) cannot be captured and annotated from the static frames, fixations and facial action units cannot be directly combined in one model due to different timescales at which labelling takes place. Video recordings in this thesis, in terms of engagement, were split into clips for the fixation measures while into frames for facial action unit recognition. One technique for a combination of two methods into one model is labelling facial action units in clips. The continuous frames are combined into one clip and that clip is labelled in terms of engagement, same as in labelling clips for the eye-tracking measure. The intensity and presence of each AU are computed by averaging the intensity and presence values respectively for all frames in this clip. However, this method will reduce the accuracy of facial AU identification for engagement, because Whitehill *et al.* [88] have demonstrated that most of the information about the facial appearance of engagement is contained in the static frames. Thus, this technique is not suitable for combining the two methods. Another technique for a combination of the two methods is labelling the raw gaze data in frames. Although previous studies have used a combination of gaze and facial action units, these studies focused on human face recognition in a static image. That is analysing the gaze direction to investigate which area is more noticeable and important on the image to help them recognise the human face. This technique is also not applicable in the context of story viewing, because the gaze direction measures the distribution of gaze data under the static image on the screen. However, characters in the digital stories (including the story-stems and movie trailers) move rapidly and the scene also changes, so gaze direction changed rapidly as the story progresses. Thus, gaze is not used in this thesis as it cannot be used to indicate which area is more noticeable and important than other areas on the screen.

Besides automatic engagement measurement methods, questionnaires were also used for collecting children's opinions for the story-stems. Chapter 4 indicated that children think they have a good understanding and attentional focus for the MCAST stories, which was

consistent with the analysis of the eye-tracking measure. However, children's answers related to the aspect of empathy showed that they could well understand the child doll's emotion represented as distress but they did not think they felt distressed as the child doll. This indicated that children's answers to the questions were not consistent with the human observation of their facial expressions. Children's unconscious facial expressions recorded by the camera could be used to analyse and reflect their mood state, but they may not express or realise their attitudes and emotions accurately towards the study by filling in a smiley-o-meter questionnaire. Therefore, the self-report measures was not the most suitable method for measuring young children's engagement levels because they may not express their attitudes and emotions accurately.

Therefore, Chapter 4 answered RQ1, where it investigated whether children's spontaneous facial expressions analysis was available to be used automatically for measuring their engagement levels in digital story-stems viewing. The facial data analysis includes eye-tracking measures and facial expressions recognition, both methods contained information related to the distinct level of child engagement. Also, the analysis of facial data showed that there was a high-level of engagement of children in this study, which suggested that digital story-stems could be used for the MCAST test and child psychiatric studies.

6.3 RQ2: How do voice type and presentation type affect child engagement levels in digital story-stems?

The important area related to engagement in this thesis is to measure children's engagement levels in digital story-stems. The story-stem approach is a reliable and valid method for investigating the important relationships in a child's life and contributing to the Attachment theory. In this method, an administrator gives the beginning of a story then asks the child to complete it, often acting out the scene using dolls. The instance of the story-stem approach used in this thesis is the *Manchester Child Attachment Story Task* (MCAST) [32].

To reduce the cost and human involvement of the story-stem approach, the way of digitalising the interaction between the child and the story administrator in the tests using story-stems was to create the interaction between the child and the computer, where the story-stems vignettes are represented on a screen. With the arrival of multimedia, the idea of merging traditional storytelling with multimedia tools is now common. Digital stories can be a mixture of graphics, animation, text, recorded audio narration, video and music, to present information on a specific topic [76]. The focus of this thesis is on designing an

engaging MCAST digital story. Chapter 5 studied two key aspects: the storytelling voice and the video format, which was answered RQ2: *How do voice type and presentation type affect child engagement levels in digital story-stems?*

Chapter 4 already demonstrated that facial data can be used to measure child engagement when discriminating high versus low engagement levels. The analysis of facial action units shows that children were engaged in digital story-stems to provide validated data for attachment assessment. Since the analysis of facial action units had a better performance than the analysis of fixations, the recognition of facial action units was used in Chapter 5 to measure children's engagement levels. A 4-class classification task was built to identify children's engagement, as answered RQ1 in Section 6.2.

Chapter 5 then investigated the two key aspects that contribute to making story-stems more engaging for children. The first part focuses on investigating the role of storytelling voice on engaging children in MCAST story-stems. Two storytellers (one male and one female) were asked to record each MCAST story-stem in two ways with different expressive qualities (expressive voice vs flat voice). A crosstabulation along with a Chi-Square analysis was conducted to test the relationship between storytelling voice types and child engagement levels using the classified frames from the 4-class classification task. The results of the Chi-Square test show that there was a significant association between storytelling voice types and child engagement levels. To analyse how the two categorical variables are related, an engagement level * voice type crosstabulation shows the frequency of different storytelling voice types broken down by child engagement level, with *column percentage* shows as the summary statistic. The *column percentage* test indicated that the percentages in columns titled 'FF' (female flat storytelling voice type) and 'ME' (male expressive storytelling voice type) categories were not significantly different according to the four child engagement levels, assigned the same subscript shown in Table 5-15. However, the percentages in columns titled 'FE' (female expressive storytelling voice type) and 'MF' (male flat storytelling voice type) categories were significantly different from each other as well as from the percentages in columns 'FF' and 'ME'. The result of the column proportions test was similar to the analyse of *post hoc* test using pitch values that found there were no significant differences between the FF and ME type according to storytelling voice types.

Across different storytelling voice types, the percentages broken down by high engagement levels (both in level 3 and 4) were compared to find the best storytelling voice type with the

highest percentage of engaged data. The crosstabulation shows that a higher percentage in both rows titled 'level 3' and 'level 4' from the story-stems narrated by a FE storytelling voice, than the percentage from the other three voice types. Since the FE voice was the best voice type for creating an engaging digital story-stem, there was no need to consider distinguishing the FF and ME voices because the performance of both were poorer than the FE voice type using the distribution of engagement classes.

Thus, Chapter 5 indicated that children were more engaged with a female expressive storytelling voice than the other voice types during the MCAST test. For the voice gender, a female voice may be a more engaging type than a male voice for children because the content of MCAST story-stems was related to "child and mother" to investigate mother-child attachment. A female voice can potentially better express a mother's emotion to help children locate themselves as the "child doll" within the story-stem. It means that children feel they are engaging and participating in that story-stem. This is essentially empathy, an important item related to engagement that could improve children's attention and comprehension in the story-stems. For the voice expressiveness of storytelling, an expressive tone of storytelling voices can more emotionally express the predicament of the MCAST story-stems, where it improves children's attention in the story to indicate a higher-level of child engagement. This was demonstrated by a higher intensity of certain action units (AU01, 02 and 04) in the expressive tone than the flat tone of storytelling voices. The combination of AU01, 02 and 04 was a reliable indicator for the facial expression of fear, which was a sign of children's feelings of anxiety and distress in the story-stems.

The second part aims to investigate the relationship between child engagement levels and two presentation types. Live-action recorded MCAST videos were taken from Chapter 4. In Chapter 5, the MCAST story-stems were redesigned using animation tools and represented on screen as animated MCAST videos, narrated by the above storytelling voices. Like analysing the storytelling voice type, a crosstabulation along with a Chi-Square analysis was also conducted to test the relationship between presentation type and child engagement level using the classified frames from the 4-class classification task. There was a significant association between child engagement levels and presentation types under a Chi-Square test. To analyse how the two categorical variables are related, an engagement level * presentation crosstabulation shows the frequency of two presentation types broken down by child engagement level, with *column percentage* shows as the summary statistic. The *column percentage* test indicated that the percentages in columns titled 'Animation' and 'Live-action'

categories were significantly different from each other according to the four child engagement levels, assigned by different subscript letters shown in Table 5-18.

Across presentation types, the percentages broken down by high engagement levels (both in level 3 and 4) were compared to find the better presentation type with a higher percentage of engaged data. The crosstabulation shows that a higher percentage in both rows titled 'level 3' and 'level 4' from the story-stems displayed with animation, than the percentage from the live-action videos. This means that both animation and live-action videos can engage children in the context of story-viewing. In addition, although live-action SAM videos are close to the real MCAST test, children were more engaged in animated MCAST videos and producing an animated MCAST video requires less time and resources, such as a camera crew and specific location. Thus, Chapter 5 suggested that the animated video was a good alternative presentation type for the design of the digital story-stems for the MCAST test.

Children's answers of the questionnaire shown in Appendix F account for their attitudes towards each of the multimedia types used for the digital story-stems in Chapter 5. The overall descriptive analysis of answers to the questionnaire were similar to results in Chapter 4, calculated by the average ratings. For example, children's answers to Q3 related to the aspect of story understanding were similar, with an average rating of 4.12 in Chapter 4 and 4.17 in Chapter 5, with 5 being the top possible rating. It indicated that all MCAST story-stems were easy to understand.

The answers were then divided into two parts to test for children's answers to each question according to the media types respectively using a Kruskal Wallis test. The result of Kruskal Wallis test indicated that there were no significant differences in children's answers to each question according to the different media types (storytelling voice and presentation). This indicates that children do not think any aspect related to their engagement would be affected when watching the story-stems redesigned by different media types. Compared to the analysis of their facial behaviours, children's answers of questionnaire were not the same as the analysis of facial action units. Children do not think their engagement level could be improved when watching the story-stems created by the media types (storytelling voice and video format) used here, while the analysis of their facial behaviours revealed that their engagement levels were different according to the watched story-stems designed by different media types. Thus, the self-report measures for children may not be the most suitable method for measuring their engagement levels because of two reasons: 1) children's answers to

questionnaire relies on their interpretation of questions and 2) children may not express their mood accurately towards the story by filling in a smiley-o-meter questionnaire. However, some of aspects in the questionnaire were still useful. For example, children had a good performance on expressing their attention and story understanding without considering the design of the digital story-stems. For studies involving children, the observed facial expressions of children would be a better engagement indicator rather than their answers to questionnaires. Therefore, the spontaneously facial expressions can be recognised as more reliable data to be used for measuring children's engagement levels in this thesis.

Therefore, Chapter 5 answered RQ2, where it focused on assessing how much each of multimedia types, including aspects of storytelling voice, and aspects of the presentation such as animation, or live-action recorded video, affected children's engagement levels. The analysis of facial action units shows that the best way of creating an engaging digital MCAST story-stem was a combination of animation and a female expressive voice. By identifying the role of different media types in digital stories on children's story experience, producing an animated story narrated by a female expressive storytelling voice was an efficient and cost-effective technique, that could be used to design other story-stems for child psychiatric studies for providing children with engaging story-watching experiences.

6.4 Limitations and Future Work

This section proposed six possible limitations in this thesis and some possible ways to overcome these limitations, as well as alternative methodologies in future work.

Sample profile

The first limitation is the sample profile. Children participants (50% males and 50% females) were recruited from several primary schools around Glasgow in Scotland. Some potentially important individual differences among children, such as their learning ability, socio-economic status and family cultures were not controlled. Children participating in the study ranged from 7 to 10 years old, there was not an equal number of children at each age.

Consequently, the study cannot measure whether children's answers of the questionnaires and their facial behaviours differed in their engaged states based on personal characteristics or their family cultures. Meanwhile, another limitation to the participant effect is no knowledge of the effect of children's gender on media types. Although the distribution of

children participated is gender-balanced, children were randomly allocated to each group in Chapter 5 so that there was no knowledge of the effect of children's gender on media types that used to create the story-stems. In future work, it will be important to assess a more homogenous sample, as well as the degree to which the results remain stable across these individual differences and across the primary school years.

Data annotation process

The second limitation concerns two aspects of the data annotation process, which was used to create the training data for machine learning. The first is coders. Although human annotation can identify and annotate the new objects in terms of engagement to make it recognisable for classification models, the accuracy of human annotation should be considered. For example, coders in this research need to annotate images in terms of engagement instead of emotion from participants' faces (i.e., a happy/sad face). This leads to annotators inconsistency revealing in that way the difficulty in annotating such kind of images. This is a generic problem of annotation. Coders' experiences of annotation highlight the need for a better understanding of the annotation process itself and cautious use of annotated data.

The second aspect is annotation window length. In this work, a fixed temporal window of 10 seconds was used to annotate the story viewing recordings. The reason was primarily to allow coding of the interactions and to simplify the machine learning application. Additionally, merging together the clips in such a way that the end of one and the start of another end up in one 10s clip ensures that all recordings can be used. The agreement of human annotation for the 'merging' clips was acceptable and it could be said that merging together the clips in this way is an available way to collect more data for measuring the engagement levels. However, fixed window length was not the only way to annotate the data. For example, if the annotation scheme was decided to apply an utterance-level segmentation, it is possible the results would have been different. In this case, the future study could be developed another annotation process that would need to have access to the boundaries of the story-stem –namely, a sentence detection may have to be applied.

Experimental materials

The limited story-stem type is a limitation. Robinson [77] indicated that researchers have embraced much greater diversity in story-stem approaches, such as MSSB [15], MCAST

[32] and Doll-play interviews [58] to inquire of the young child about how they think and feel about important relationships. Common to all of the story stem approaches is that they seek to engage the child in the story vignette to respond to a challenging situation. However, in this research, only one instance (MCAST) of the story-stem approach was used and this means that researchers currently know relatively little about if the work extends to other story types. Thus, examining how engaged children use different story-stems is critical. Future research could, for example, test more story-stems to see if the result of this research hold.

Conditions

Chapter 5 investigated the effect of two media types, the voice and presentation style, on engaging children in the digital story-stem approach. Firstly, the use of storytelling voice is subject to several limitations. When investigating the storytelling voice in this research, only two storytellers (one actor and actress) were recruited to narrate the MCAST story-stems. Future work could focus on training more storytellers (both males and females) to present the story-stems to analyse the effects of voice gender on children's engagement levels to see how the results generalise. Meanwhile, since no computer-generated voices that can currently imitate the dynamic, expressive range of the voices of fully trained MCAST administrators, future work could also focus on investigating the differences between computer-generated voices and human storytelling voices. That could help researchers create an engaging computer-generated voice used for digital story-stems, which have a further reduction of cost and human involvement for the tests using the story-stem approach.

Secondly, the design of using different presentation types is also subject to limitations. Although children were more engaged in animated story-stems that were redesigned using animation tools in this research, the symbolic screen dolls in the current version of the animated story-stems were not the same as the physical dolls (see Figure 1-2, the SAM's setup). The effects of this are unknown, so future work could investigate two things: 1) redesigning a new version of animated story-stems that used models of the physical dolls; 2) detecting whether the difference design of dolls affects children to complete the story related to their Attachment status.

Lastly, besides the storytelling voices and presentation types used in this thesis, other media types, such as music, could also be investigated to study their role on children's engagement for designing an engaging story-stems in the future.

Classification and Sample size

The results reported herein also should be considered in the light of some limitations. Chapter 4 indicated that the performance of the classifier has a low accuracy of classifying frames/clips for the not-engaged and rarely-engaged categories. Also, Chapter 5 indicated that the 4-class classification task was less reliable than the binary classification in Chapter 4, as the ability to identify frames with the rarely-engaged category was poorer than the other three categories. One reason for this lower accuracy was the small sample size in these low engagement categories so that statistical tests would not be able to identify significant relationships within data set. Future work could consider the sample size of the study, such as enlarging the sample size. Basing the study in larger sample size could have generated more accurate results.

Equipment

The equipment used in this research was not only focused on the engagement measurement. Finally, the test using the story-stem approach includes 1) bringing a child into a deep engagement; 2) asking the child to complete the story by playing with dolls. During the SAM test, the child was given the instruction to “press the button to go to the next story” after he/she completed this story. When a child was given that instruction, the child sometimes reached forward to select it on the screen using their finger instead of the physical button on the desk. This behaviour may be natural for children to select the button using a finger. Future work could, for example, introduce a touchscreen to the SAM system or other computerised MCAST system so that the child no longer needs to interact with the system using a mouse or a physical button, and can use a technique that is natural to them.

6.5 Conclusion

This thesis focuses on measuring children’s engagement in digital story-stems, specifically on using one instance of the story-stem approach: MCAST. Automated MCAST assessments need to present story-stems in a cost-effective way on a laptop screen to digitalise the interaction between the child and the story, without disrupting the storytelling. However, providing such tests via a computer relies on the child being engaged in the digital story. If they are not engaged, then the tests will not be successful and the collected data will be of poor-quality, which will not allow for the MCAST assessment. Automated measures were used to create a tool to identify children’s engagement levels from a video recording of their facial expressions when watching the story-stems. Meanwhile, bringing a child into

a deep engagement while watching the digital story-stems vignettes could reduce the chance of poor-quality data assessment. This thesis is focused on investigating the effects of multimedia tools for creating a better and more engaging digital story to make children more easily get engaged to complete the test in spontaneous play. Therefore, there are two main contributions in this thesis corresponding to the research questions:

- 1) Children's spontaneous facial expressions can be used to automatically measure their engagement levels in digital story-stems;
- 2) Both presentation type (video formats) and voice type (storytelling voices) affect child engagement levels in digital stories. The best way of creating an engaging MCAST story was using animations to display by the movement of two symbolic screen 'dolls', narrated with a female expressive storytelling voice.

The measurement procedure of child engagement can be implemented using spontaneous facial data with low cost algorithms and in a non-invasive way with simple sensors. It reduced the need for so much time from trained assessors and ensures the quality of the data that will be used to make assessments. Researchers could acquire large amounts of facial data to measure child engagement levels automatically to improve the efficiency of coding evaluations.

As children were engaged in the digital story-stems in this thesis, the digital story-stem approach can be recognised as a reliable and cost-effective method for the MCAST test and other child psychiatric studies that used the story-stems. This is the first attempt at automating Attachment administration using the digital story-stem approach. The use of digital story-stems was able to digitalise the interaction between children and the storyteller so and reduce the cost and human involvement of the child psychiatry tests. Meanwhile, when displaying digital story-stems to children, people without MCAST training, such as teachers, could also administer the MCAST test. This means that more children will be tested.

Moreover, an engaging digital story-stems represented on a screen could bring a child into a deep engagement to reduce the chance of poor-quality data assessment. The best engaging story-stems created in this thesis, design in animated video narrated with a female expressive storytelling voice, could be used for the computerised MCAST systems (e.g., the SAM system [78]). Specifically, when children were watching the best engaging story-stems, they were happier to complete the MCAST story for the test. The spontaneously completed story

and children's behaviours can be recognised as more reliable data to be used for evaluating children's attachment status, which is the final purpose of the MCAST test.

Automated engagement measurement and the use of the digital story-stem approach could improve the efficiency of Attachment assessments such as MCAST. Automating child Attachment assessment has the potential to screen Attachment across the population and identify children with disorganised family attachment that need attention. We believe that our research will significantly improve population health and wellbeing.

Bibliography

- [1] Ainsworth, S. 2008. How do animations influence learning? *Current Perspectives on Cognition, Learning, and Instruction: Recent Innovations in Educational Technology that Facilitate Student Learning*. (2008), 37–67.
- [2] Alvarado, J. et al. 2008. Network News - engagement in classroom. *Preschool Network*. 6, 4 (2008).
- [3] Anzalone, S.M. et al. 2015. Evaluating the Engagement with Social Robots. *International Journal of Social Robotics*. 7, 4 (2015), 465–478.
DOI:<https://doi.org/10.1007/s12369-015-0298-7>.
- [4] Appleton, J.J. et al. 2006. Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*. 44, 5 (2006), 427–445. DOI:<https://doi.org/10.1016/j.jsp.2006.04.002>.
- [5] Arnett 1989. Caregiver Interaction Scale: Evaluation Tool. (1989).
- [6] Attfield, S. et al. 2011. Towards a science of user engagement (Position Paper). *WSDM Workshop on User Modelling for Web Applications*. (2011).
- [7] Bal, E. et al. 2010. Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *Journal of Autism and Developmental Disorders*. 40, 3 (2010), 358–370. DOI:<https://doi.org/10.1007/s10803-009-0884-3>.
- [8] Baltrusaitis, T. et al. 2015. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. (2015), 1–6.
DOI:<https://doi.org/10.1109/FG.2015.7284869>.
- [9] Baltrusaitis, T. et al. 2016. OpenFace: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. (2016). DOI:<https://doi.org/10.1109/WACV.2016.7477553>.
- [10] Bednarik, R. et al. 2012. Gaze and Conversational Engagement in Multiparty Video Conversation: An Annotation Scheme and Classification of High and Low Levels of Engagement. *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. (2012), 10:1-10:6.
DOI:<https://doi.org/10.1145/2401836.2401846>.
- [11] Bickmore, T. et al. 2010. Maintaining Engagement in Long-term Interventions with Relational Agents. *Appl Artif Intell*. 24, 6 (2010), 648–666.
DOI:<https://doi.org/10.1080/08839514.2010.492259>.
- [12] Bidwell, J. and Fuchs, H. 2011. Classroom Analytics: Measuring Student Engagement with Automated Gaze Tracking. November 2011 (2011), 1--17.
DOI:<https://doi.org/10.13140/RG.2.1.4865.6242>.
- [13] Blom, P.M. et al. 2014. Towards personalised gaming via facial expression recognition. *Proceedings of Tenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. AIIDE (2014), 30–36.
- [14] Bretherton, I. et al. 1990. Family relationships as represented in a story-completion task at thirty-seven and fifty-four months of age. *New Directions for Child and Adolescent Development*. (1990). DOI:<https://doi.org/10.1002/cd.23219904807>.
- [15] Bretherton, I. et al. 2003. The MacArthur Story Stem Battery and Parent-Child Narratives. *Revealing the Inner Worlds of Young Children: The MacArthur Story Stem Battery and Parent-Child Narratives*.

- [16] Busselle, R. and Bilandzic, H. 2008. Fictionality and perceived realism in experiencing stories: A model of narrative comprehension and engagement. *Communication Theory*. 18, 2 (2008), 255–280. DOI:<https://doi.org/10.1111/j.1468-2885.2008.00322.x>.
- [17] Busselle, R. and Bilandzic, H. 2009. Measuring narrative engagement. *Media Psychology*. 12, 4 (2009), 321–347. DOI:<https://doi.org/10.1080/15213260903287259>.
- [18] Carl, B. 2010. *Child Caregiver Interaction Scale (CCIS) Revised Edition*.
- [19] Chang, C.-C. and Lin, C.-J. 2011. LIBSVM. *ACM Transactions on Intelligent Systems and Technology*. (2011). DOI:<https://doi.org/10.1145/1961189.1961199>.
- [20] Cortes, C. and Vapnik, V. 1995. Support-Vector Networks. *Machine Learning*. (1995). DOI:<https://doi.org/10.1023/A:1022627411411>.
- [21] Cutrell, E. and Way, M. 2007. What Are You Looking For? An Eye-tracking Study of Information Usage in Web Search. (2007), 407–416.
- [22] Davis 2004. What does it Mean for Students to Be Engaged? *SAGE Publication*. (2004), 21–33. DOI:<https://doi.org/http://dx.doi.org/10.4135/9781483387383.n2>.
- [23] Doherty, K. and Doherty, G. 2018. Engagement in HCI: Conception, Theory and Measurement. *ACM Computing Surveys*. 51, 5 (2018), 99:1–39. DOI:<https://doi.org/10.1145/3234149>.
- [24] Ekman, P. 2010. Symposium on Emotion New Findings , New Questions. 3, 1 (2010), 34–38.
- [25] Ekman, P. and Friesen, W. V. 1978. Facial Action Coding System: Manual. (1978), 233.
- [26] Falck-Ytter, T. et al. 2015. Eye Contact Modulates Cognitive Processing Differently in Children With Autism. *Child Development*. 86, 1 (2015), 37–47. DOI:<https://doi.org/10.1111/cdev.12273>.
- [27] Finn, J.D. 1989. Withdrawing from School. *Review of Educational Research*. 59, 2 (1989), 117. DOI:<https://doi.org/10.2307/1170412>.
- [28] Fredricks, J.A. 2011. Engagement in school and out-of-school contexts: A multidimensional view of engagement. *Theory into Practice*. 50, 4 (2011), 327–335. DOI:<https://doi.org/10.1080/00405841.2011.607401>.
- [29] Fredricks, J.A. et al. 2004. School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*. 74, 1 (2004), 59–109. DOI:<https://doi.org/10.3102/00346543074001059>.
- [30] Goldwyn, R. et al. 2000. The Manchester Child Attachment Story Task: relationship with parental AAI, SAT and child behaviour. *Attachment & human development*. 2, 1 (2000), 71–84. DOI:<https://doi.org/10.1080/146167300361327>.
- [31] Grafsgaard, J.F. et al. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. *International Conference on Educational Data Mining*. (2013).
- [32] Green, J. et al. 2000. A new method of evaluating attachment representations in young school-age children : The Manchester Child Attachment Story Task. *Attachment & Human Development*. 2, 1 (2000), 48–70. DOI:<https://doi.org/10.1080/146167300361318>.
- [33] Griffiths, A.J. et al. 2009. School, Student Engagement and Positive Adaptation. *Handbook of Positive Psychology in the Schools*. M.. Furlong et al., eds. 197–211.
- [34] Hanna, L. et al. 2004. Evaluating Computer Game Concepts with Children. *IDC '04 Proceedings of the 2004*. (2004), 49–56. DOI:<https://doi.org/10.1145/1017833.1017840>.
- [35] Harrigan, J.A. and O'Connell, D.M. 1996. How do you look when feeling anxious? Facial displays of anxiety. *Personality and Individual Differences*. 21, 2 (1996), 205–212. DOI:[https://doi.org/10.1016/0191-8869\(96\)00050-5](https://doi.org/10.1016/0191-8869(96)00050-5).

- [36] Hernandez, J. et al. 2013. Measuring the engagement level of TV viewers. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*. (2013). DOI:<https://doi.org/10.1109/FG.2013.6553742>.
- [37] Hong, W. et al. 2004. Does Animation Attract Online Users' Attention? The Effects of Flash on Information Search Performance and Percept. *Information Systems Research*. 15, 1 (2004), 60–86. DOI:<https://doi.org/10.1287/isre.1040.0017>.
- [38] VAN IJZENDOORN, M.H. et al. 1999. Disorganized attachment in early childhood: Meta-analysis of precursors, concomitants, and sequelae. *Development and Psychopathology*. (1999). DOI:<https://doi.org/10.1017/s0954579499002035>.
- [39] Ishii, R. et al. 2011. Combining Multiple Types of Eye-gaze Information to Predict User ' s Conversational Engagement. *Human Factors*. (2011), 1–8.
- [40] Ishii, R. et al. 2013. Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems*. 3, 2 (2013), 11:1-11:25. DOI:<https://doi.org/10.1145/2499474.2499480>.
- [41] Ishii, R. and Nakano, Y.I. 2008. Estimating user's conversational engagement based on gaze behaviors. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 5208 LNAI, (2008), 200–207. DOI:https://doi.org/10.1007/978-3-540-85483-8_20.
- [42] Ishii, R. and Yukiko, I.N. 2010. An Empirical Study of Eye-gaze Behaviors : Towards the Estimation of Conversational Engagement in Human-Agent Communication. *EGIHMI '10 Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction*. (2010), 33–40. DOI:<https://doi.org/10.1145/2002333.2002339>.
- [43] Karimi, A. and Lim, Y.P. 2010. Children, engagement and enjoyment in digital narrative. *ASCILITE 2010 - The Australasian Society for Computers in Learning in Tertiary Education*. (2010), 475–483.
- [44] Kiliánová, G. 1999. Women ' S and Men ' S Storytelling : What Is the Difference ? Some Observations in Contemporary Slovak. 5, (1999), 99–108.
- [45] Kory Westlund, J.M. et al. 2017. Flat vs. Expressive Storytelling: Young Children's Learning and Retention of a Social Robot's Narrative. *Frontiers in Human Neuroscience*. 11, June (2017), 1–20. DOI:<https://doi.org/10.3389/fnhum.2017.00295>.
- [46] De Kruif, R.E.L. et al. 2000. Classification of teachers' interaction behaviors in early childhood classrooms. *Early Childhood Research Quarterly*. 15, 2 (2000), 247–268. DOI:[https://doi.org/10.1016/S0885-2006\(00\)00051-X](https://doi.org/10.1016/S0885-2006(00)00051-X).
- [47] Lee, D. et al. 2015. Measuring the engagement level of children for multiple intelligence test using Kinect. *Proc. SPIE 9445, Seventh International Conference on Machine Vision (ICMV 2014)* (2015).
- [48] Leite, I. et al. 2015. Comparing Models of Disengagement in Individual and Group Interactions. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15* (2015), 99–105.
- [49] Lim, K. et al. 2010. Measuring attachment in large populations: A systematic review. *Educational and Child Psychology*. (2010).
- [50] Littlewort, G.C. et al. 2011. Automated measurement of children ' s facial expressions during problem solving tasks. *IEEE International Conference on. IEEE*. (2011), 30–35. DOI:<https://doi.org/10.1109/FG.2011.5771418>.
- [51] Marshall, K. et al. 2015. Supporting Children to Engage in Play for Wellbeing. *CHI'15 Extended Abstracts* (Seoul, Republic of Korea, 2015), 2445–2448.

- [52] McWilliam, R.A. et al. 2003. Adult Interactions and Child Engagement. *Early Education & Development*. 14, 1 (2003), 7–28. DOI:https://doi.org/10.1207/s15566935eed1401.
- [53] McWilliam, R.A. and Bailey, D.B. 1995. Effects of Classroom Social Structure and Disability on Engagement. *Topics in Early Childhood Special Education*. 15, 2 (1995), 123–147. DOI:https://doi.org/10.1177/027112149501500201.
- [54] Miller, B.W. 2015. Using reading times and eye-movements to measure cognitive engagement. *Educational Psychologist*. 50, March (2015), 31–42. DOI:https://doi.org/10.1080/00461520.2015.1004068.
- [55] MINNIS, H. et al. 2010. The Computerised Manchester Child Attachment Story Task: a novel medium mpr_324 233..242 for assessing attachment patterns. *International Journal of Methods in Psychiatric Research*. 19, 4 (2010), 233–242. DOI:https://doi.org/10.1002/mpr.
- [56] Monkaresi, H. et al. 2017. Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*. 8, 1 (2017), 15–28. DOI:https://doi.org/10.1109/TAFFC.2016.2515084.
- [57] Munn, S.M. and Pelz, J.B. 2008. Fixation-identification in dynamic scenes : Comparing an automated algorithm to manual coding. *Science*. 1, 212 (2008), 33–42.
- [58] Murray, L. et al. 1999. Children’s social representations in dolls’ house play and theory of mind tasks, and their relation to family adversity and child disturbance. *Social Development*. (1999). DOI:https://doi.org/10.1111/1467-9507.00090.
- [59] Nakano, I. and Ishii, R. 2010. Estimating User ’ s Engagement from Eye-gaze Behaviors in Human-Agent Conversations. (2010).
- [60] National Research Council and Institute of Medicine 2004. *Engaging Schools: Fostering High School Students’ Motivation to Learn*. The National Academies Press.
- [61] O’Brien, H.L. et al. 2009. Developing and evaluating a reliable measure of user engagement. *Proceedings of the American Society for Information Science and Technology* (2009), 1–10.
- [62] O’Brien, H.L. et al. 2016. Investigating the Role of User Engagement in Digital Reading Environments. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval - CHIIR ’16*. (2016), 71–80. DOI:https://doi.org/10.1145/2854946.2854973.
- [63] O’Brien, H.L. and Toms, E.G. 2010. The Development and Evaluation of a Survey to Measure User Engagemen. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*. 61, 1 (2010), 50–69. DOI:https://doi.org/10.1002/asi.
- [64] O’Brien, H.L. and Toms, E.G. 2008. What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*. 59, 6 (2008), 938–955. DOI:https://doi.org/10.1002/asi.
- [65] Olsen, A. 2012. *Determining the Tobii I-VT Fixation Filter ’ s Default Values*.
- [66] Olsen, A. 2012. *The Tobii I-VT Fixation Filter: Algorithm description*.
- [67] Page, T. and Bretherton, I. 2001. Mother- and father-child attachment themes in the story completions of pre-schoolers from post-divorce families: do they predict relationships with peers and teachers? *Attachment & human development*. 3, 1 (2001), 1–29. DOI:https://doi.org/10.1080/713761897.
- [68] Patil, P.R. and Manjare, C.A. 2014. Expressive speech analysis for story telling application. *2014 IEEE Global Conference on Wireless Computing & Networking (GCWCN)*. (2014), 97–101. DOI:https://doi.org/10.1109/GCWCN.2014.7030856.

- [69] Pernet, C.R. and Belin, P. 2012. The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*. 3, (2012), 1–11. DOI:https://doi.org/10.3389/fpsyg.2012.00023.
- [70] Peters, C. et al. 2009. An exploration of user engagement in HCI. *Proceedings of the International Workshop on Affective Aware Virtual Agents and Social Robots*. (2009), 1–3. DOI:https://doi.org/10.1145/1655260.1655269.
- [71] Poole, A. and Ball, L.J. 2005. Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. *Encyclopedia of Human-Computer Interaction*. (2005), 211–219. DOI:https://doi.org/10.4018/978-1-59140-562-7.
- [72] Raspa, M.J. et al. 2001. Child Care Quality and Children's Engagement. *Early Education and Development*. 12, 2 (2001), 209–224. DOI:https://doi.org/10.1207/s15566935eed1202.
- [73] Read, J. et al. 2002. Endurability, Engagement and Expectations: Measuring Children's Fun. *Interaction Design and Children*. 2, (2002), 1–23. DOI:https://doi.org/10.1.1.100.9319.
- [74] Reeve, J. and Tseng, C.M. 2011. Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology*. 36, 4 (2011), 257–267. DOI:https://doi.org/10.1016/j.cedpsych.2011.05.002.
- [75] Rich, C. et al. 2010. Recognizing engagement in human-robot interaction. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. (2010), 375–382. DOI:https://doi.org/10.1109/HRI.2010.5453163.
- [76] Robin, B.R. 2006. The educational uses of digital storytelling. *Society for Information Technology & Teacher Education International Conference*. (2006), 709–716. DOI:https://doi.org/10.1016/j.sbspro.2012.11.424.
- [77] Robinson, J.L. 2007. Story stem narratives with young children: moving to clinical research and practice. *Attachment & human development*. 9, 3 (2007), 179–185. DOI:https://doi.org/10.1080/14616730701453697.
- [78] Roffo, G. et al. 2019. Automating the administration and analysis of psychiatric tests: The case of attachment in school age children. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK, 2019), 595:1-595:12.
- [79] Salvucci, D.D. and Goldberg, J.H. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. *Proceedings of the Eye Tracking Research and Applications Symposium*. (2000), 71–78. DOI:https://doi.org/10.1145/355017.355028.
- [80] Schiavo, G. et al. 2014. Engagement recognition using easily detectable behavioral cues. *Intelligenza Artificiale*. 8, May 2016 (2014), 197–210. DOI:https://doi.org/10.3233/IA-140073.
- [81] Sidner, C.L. et al. 2004. Where to look: a study of human-robot engagement. *Proceedings of the 9th international conference on Intelligent User Interfaces* (2004), 78–84.
- [82] Sidner, C.L. and Dzikovska, M. 2002. Human - Robot Interaction: Engagement between Humans and Robots for Hosting Activities. *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)* (2002).
- [83] Sinatra, G.M. et al. 2015. The Challenges of Defining and Measuring Student Engagement in Science. *Educational Psychologist*. 50, 1 (2015), 1–13. DOI:https://doi.org/10.1080/00461520.2014.1002924.
- [84] Theune, M. et al. 2006. Generating expressive speech for story telling applications. *IEEE Transactions on Audio, Speech, and Language Processing*. 14, 4 (2006), 1099–1108.
- [85] Titze, I.R. 1994. *Principles of Voice Production*. Englewood Cliffs: Prentice Hall.
- [86] University of Colorado Digital Storytelling Assignment: Rubric Example. 1.

- [87] Vo, D.-B. et al. 2017. SAM : The School Attachment Monitor. *IDC'17* (Stanford, CA, USA, 2017), 671–674.
- [88] Whitehill, J. et al. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*. 5, 1 (2014), 86–98. DOI:<https://doi.org/10.1109/TAFFC.2014.2316163>.
- [89] Wiebe, E.N. et al. 2014. Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior*. 32, (2014), 123–132. DOI:<https://doi.org/10.1016/j.chb.2013.12.001>.
- [90] Wotschack, C. 2009. *Eye Movements in Reading Strategies: How Reading Strategies Modulate Effects of Distributed Processing and Oculomotor Control*.
- [91] Xie, L. et al. 2008. Are tangibles more fun? *Proceedings of the 2nd international conference on Tangible and embedded interaction - TEI '08*. (2008), 191–198. DOI:<https://doi.org/10.1145/1347390.1347433>.
- [92] Xu, Q. et al. 2013. Designing engagement-aware agents for multiparty conversations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. (2013), 2233–2242. DOI:<https://doi.org/10.1145/2470654.2481308>.
- [93] Yu, C. et al. 2004. Detecting User Engagement in Everyday Conversations. (2004). DOI:<https://doi.org/10.1152/japplphysiol.00536.2007>.
- [94] Yun, W.-H. et al. 2015. Automatic Engagement Level Estimation of Kids in a Learning Environment. *International Journal of Machine Learning and Computing*. 5, 2 (2015), 148–152. DOI:<https://doi.org/10.7763/IJMLC.2015.V5.499>.
- [95] Yun, W.-H. et al. 2018. Automatic Recognition of Children Engagement from Facial Video using Convolutional Neural Networks. *IEEE Transactions on Affective Computing*. 3045, c (2018), 1–1. DOI:<https://doi.org/10.1109/TAFFC.2018.2834350>.

Appendices

Appendix A The User Engagement Scale (UES)

-
1. I lost myself in this shopping experience.
 2. I was so involved in my shopping task that I lost track of time.
 3. I blocked out things around me when I was shopping on this website.
 4. When I was shopping, I lost track of the world around me.
 5. The time I spent shopping just slipped away.
 6. I was absorbed in my shopping task.
 7. During this shopping experience I let myself go.
 8. I was really drawn into my shopping task.
 9. I felt involved in this shopping task.
 10. This shopping experience was fun.
 11. I continued to shop on this website out of curiosity.
 12. The content of the shopping website incited my curiosity.
 13. I felt interested in my shopping task.
 14. Shopping on this website was worthwhile.
 15. I consider my shopping experience a success.
 16. This shopping experience did not work out the way I had planned.*
 17. My shopping experience was rewarding.
 18. I would recommend shopping on this website to my friends and family.
 19. This shopping website is attractive.
 20. This shopping website was aesthetically appealing.
 21. I liked the graphics and images used on this shopping website.
 22. This shopping website appealed to my visual senses.
 23. The screen layout of this shopping website was visually pleasing.
 24. I felt frustrated while visiting this shopping website.*
 25. I found this shopping website confusing to use.*
 26. I felt annoyed while visiting this shopping website.*
 27. I felt discouraged while shopping on this website.*
 28. Using this shopping website was mentally taxing.*
 29. This shopping experience was demanding.*
 30. I felt in control of my shopping experience.
 31. I could not do some of the things I needed to do on this shopping website.*
-

The scale was administered using a five-point scale with “strongly disagree” and “strongly agree” at the respective endpoints. Items identified with an asterisk (*) indicate items that were reverse-coded.

Appendix B Items used for developing the narrative engagement scale [17]

Empathy

- EP1*: At key moments in the film, I felt I knew exactly what the characters were going through emotionally. (adapted from Cohen, 2001)
- EP2: At important moments in the film, I could feel the emotions the characters felt. (adapted from Cohen, 2001)
- EP3: During the program, when a main character succeeded, I felt happy, and when they suffered in some way, I felt sad. (adapted from Cohen, 2001)
- EP4: I never really shared the emotions of the characters (–).
- EP5: The story affected me emotionally. (T; Green & Brock, 2000)
-

Sympathy

- S1: I felt sorry for some of the characters in the program.
- S2: I was embarrassed for some of the characters in the program.
- S3: I was worried for some of the characters in the program.
-

Cognitive perspective taking

- CP1: I was able to understand the events in the program in a way similar to the way the characters understood them. (adapted from Cohen, 2001)
- CP2: I understood the reasons why the characters did what they did. (adapted from Cohen, 2001)
- CP3: I could understand why the characters felt the way they felt.
- CP4: My understanding of the characters is unclear. (–)(adapted, Cohen, 2001)
- CP5: It was difficult to understand why the characters reacted to situations as they did. (–)
- CP6: I could easily imagine myself in the situation of some of the characters. (adapted from Cohen, 2001)
-

Loss of time

- LT1: During the program, I lost track of time.
- LT2: The program seemed to drag. (–)
- LT3: When the program ended, I was surprised that it was over so quickly.
-

Loss of self-awareness

- LS1: At times during the program, I completely forgot that I was in the middle of an experiment.
- LS2: I forgot my own problems and concerns during the program.
- LS3: While watching, I found myself thinking about what I had done before the experiment or what I would do after it. (–)
-

Narrative presence

- NP1: At times during the program, the story world was closer to me than the real world. (adapted from Kim & Biocca, 1997)
- NP2: My attention was focused more on my surroundings than on the program. (–)
- NP3: The program created a new world, and then that world suddenly disappeared when the program ended. (adapted from Kim & Biocca, 1997)
- NP4: During the program, my body was in the room, but my mind was inside the world created by the story. (adapted from Kim & Biocca, 1997)
-

Narrative involvement

NI1: I was mentally involved in the story while viewing. (T; Green & Brock, 2000)

NI2: I was never really pulled into the story. (–)

NI3: While viewing I was completely immersed in the story. (Appel et al., 2002)

NI4: Overall, the viewing experience was intense for me. (Appel et al., 2002)

NI5: I wanted to learn how the story ended. (T; Green & Brock, 2000)

NI6: While viewing I wanted to know how the events would unfold. (Appel et al., 2002)

Distraction

D1: I found my mind wandering while the program was on. (–) (T; Green & Brock, 2000)

D2: While the program was on I found myself thinking about other things. (–) (Appel et al., 2002)

D3: I had a hard time keeping my mind on the program. (–)

Ease of cognitive access

EC1: I could easily follow the action and events. (Appel et al., 2002)

EC2: I had a hard time recognizing the thread of the story. (–) (Appel et al., 2002)

EC3: I had to work to stay focused on the story. (–) (Appel et al., 2002)

Narrative realism

NR1: The story was logical and convincing.

NR2: I understood why the events unfolded the way they did.

NR3: At some points in the story, it was not quite clear why something happened. (–)

NR4: At points, I had a hard time making sense of what was going on in the program. (–)

Additional transportation items (Green & Brock, 2000)

While I was watching the movie, I could easily picture the events in it taking place.

I could picture myself in the scene of the events shown in the movie.

After finishing the movie, I found it easy to put it out of my mind. (–)

I found myself thinking of ways the story could have turned out differently.

The events in the story are relevant to my everyday life.

The events in the story have changed my life.

While I was watching the movie, activity going on in the room around me was on my mind. (–)

Note: Items belonging to the transportation scale are marked with "T" if they belong to one of the dimensions of the narrative engagement scale. Items marked with (–) were reversed coded.

*Key to items' original theoretical constructs: CP = cognitive perspective taking; EP = empathy; SM = sympathy; NP = narrative presence; NI = narrative involvement; LT = loss of time; LS = loss of self; EC = ease of cognitive access; DS = distraction; NR = narrative realism.

Appendix C Definitions of Levels and Types of the E- Qual Coding System

CODES	DEFINITIONS
LEVELS	
Persistence ^a	Involves some problem-solving and some challenge; includes either changing strategies or using the same strategy; requires goal-directed behavior. <i>Example:</i> A child is trying to complete a puzzle, but is having difficulty fitting the pieces together. She tries several different ways to make the two fit.
Symbolic ^a	Involves use of conventional forms of behavior such as language, pretend play, sign language, drawings, etc.; requires decontextualization. <i>Example:</i> A child is in the dramatic play area and is pretending to cook dinner.
Encoded ^a	Involves use of conventional forms of behavior that are context bound and that depend on referents; includes use of understandable language. <i>Example:</i> A child is talking to the teacher about the game they are playing.
Constructive ^a	Involves manipulating objects to create or build something; includes putting objects together in some type of spatial form; requires intentionality. <i>Example:</i> A child is drawing a picture, or building a tower.
Differentiated ^b	Involves coordination and regulation of behavior that reflects elaboration and progress toward conventionalization; includes active interaction with environment. <i>Example:</i> A child is using a spoon to eat during mealtime.
Focused Attention ^c	Involves watching or listening to features in the environment for a duration of at least 3 seconds; includes physical characteristics such as serious facial expression and subdued motor activity. <i>Example:</i> A child watches the teacher while she reads a book during circle time.
Undifferentiated ^d	Involves interacting with the environment without differentiating behavior; includes repetitive, low-level behaviors. <i>Example:</i> A child bangs two blocks together over and over.
Casual Attention ^d	Involves relaxed and wide-ranging attention; includes monitoring of the environment. <i>Example:</i> A child looks around the room to find the teacher.
Nonengaged ^d	Involves lack of occupation; requires the absence of any of the other behaviors. <i>Example:</i> A child pushes another child while standing in line; a child sits and stares off into space during center time.
TYPES	
Peers	Displays one of the above behaviors towards or with peers.
Adults	Displays one of the above behaviors towards or with adults.
Objects	Displays one of the above behaviors towards or with toys, materials, or other aspects of the physical environment.
Self	Displays one of the above behaviors towards or with himself or herself.

^a Sophisticated engagement.

^b Differentiated engagement.

^c Focused engagement.

^d Unsophisticated engagement.

Appendix D Information pack to children's family

a) Opt-out Consent Form



University of Glasgow | College of Science & Engineering

Reply slip to opt out of this project

Project Title: Evaluating child engagement in short story-stems taken from Manchester Child Attachment Story Task (MCAST)

PLEASE WRITE YOUR NAME BELOW TO SAY YOU **DO NOT** WISH YOUR CHILD TO TAKE PART IN THIS STUDY

Your name (parent/guardian) _____ Date _____

Name of child's name _____

Your signature _____ Date _____

We will be very grateful if you could return this slip to us in the pre-paid envelope enclosed as soon as possible and no later than **Friday 26th September**, as we plan to start our experiment shortly after that.

Please feel free to contact me at the phone number/email if you have any questions about your child taking part. Experimenter Details:

Miss Rui Huan

Email: r.huan.1@research.gla.ac.uk Phone: +44 7746 8874 77

Office: F122, School of Computing Science, University of Glasgow, 18 Lilybank Garden, Glasgow, G12 8RZ

This study has been approved by the Ethics Committee (Reference Number –300150184).

b) Participant Information Sheet/ Letter to Child



What children need to know about our study?

PARENTS! Please read through this information sheet with your child.

Project Title: Evaluating child engagement in short story-stems taken from Manchester Child Attachment Story Task (MCAST)

We want to ask you if you would like to take part in a research project which we think you may enjoy. You can talk to anyone about this – for example your family or your teacher. We will do our best to give you any information that you may want. You do not have to decide now.

We are designing a method of measuring child engagement by using some psychological story-stems displayed in short videos. These story stems are taken from Manchester Child Attachment Story Task (MCAST)*.



*MCAST is a structured doll play method for examining children's feeling about their family relationship and we will use three story stems from MCAST.

We are asking 7-10 years old children to try out our experiment and tell us what they think about these videos. Eventually, we hope that this experiment will tell us about how to improve child engagement in

a special circumstance. But, for now, all we want to do is work with Glasgow children to estimate their engagement.



If you decide to take part, you would spend about half an hour trying out the experiment. You would listen to some stories told to you by a

computer. And you would be asked to finish the story. At the end, you would tell us your feeling about each story by filling a questionnaire. This would be videotaped and your eye movement would also be recorded by using a desktop eye tracker.

The video tapes will ONLY be viewed by the researchers. We will look carefully at the videotapes of all the children who take part so that we can learn how to make our research better.

You don't have to take part. Not everyone who volunteers will be able to do our study. It depends on how many people wish to take part. We plan to involve 20 children across primary schools in Glasgow City.

Thank you for taking the time to read this information sheet and for considering taking part in this study.

Miss Rui Huan, experimenter of this project, University of Glasgow

c) Letter to Parents/Carers



Project Title: Evaluating child engagement in short story-stems taken from Manchester Child Attachment Story Task (MCAST)

Dear Parents and Carers,

Your child is being invited to take part in a research study to estimate child engagement in some psychological story stems displayed in short videos. We prepare two information sheets for you and your child respectively. Before you talk to your child about this research, it is important for you to understand why this research is being done and what it will involve. Please take time to read the following information carefully and do not hesitate to ask the experimenter if there is anything you do not understand or if you would like to know further information. Thank you for reading this.

What is the purpose of the study?

We propose a method of estimating child engagement in some psychological story stems displayed in short videos. In particular, we present a statistical analysis of children's eye movements by using a desktop eye-tracker to measure if they are engaged in these videos.

What is involved in the study?

A warm-up explanation represents an introduction to the procedure. Then videos of the 'distress' vignettes will be displayed to your child by using a PC screen. There are four videos with different voice conditions (i.e. actor's boring voice, actor's exciting voice, actress's boring voice and actress's exciting voice) in each vignette. Participants need to watch 12 videos in total. These psychological story stems are taken from the standard Manchester Child Attachment Story Task (MCAST), which is a method for assessing mental health.

For each vignette, there are something stressful represented in the child doll. In this phase, your child would be engaged and your information will be collected. There are two 5-minute breaks after the fourth and eighth video respectively. After watching videos, participants will be asked to complete a demographic questionnaire.

The experiment takes no more than 1 hour to administer. It will be carried out in the School of Computing Science, University of Glasgow. The address is Sir Alwyn Williams Building, 18 Lilybank Gardens, Glasgow, UK, G12 8RZ.

Does my child have to take part in this study?

No, their participation in this project is entirely voluntary and they are free to withdraw at any time without explanation.

In addition, for this study, the University of Glasgow Research Ethics Committee have allowed us to take a consent form; we ask you to reply to us in the attached slip **if you wish your child to take**

part. If you are happy for your child to take part, please read the participant information sheet with your child.

What type of information will be sought from my child? Will my child be recorded, and how will the recorded media be used?

During the experiment, children eye movement will be recorded by using a Tobii EyeX eye-tracker. Children's activities in this experiment will also recorded by audio and/or video recordings with your permission.

There is a questionnaire that your child needs to complete and you child's name will not be recorded on the questionnaire and all information will not be disclosed to other parties.

Will my child taking part in this project be kept confidential?

All information which is collected about your child, or responses that your child provides, during the course of the research will be kept strictly confidential. Your child will be identified by an ID number and any information about your child will have name and address removed so that your child cannot be recognised from it.

All the electronic data files generated by our experiment (i.e. children's eye movement data as well as audio and/or video recording) will be removed from the laptop and move to an encrypted external hard drive. The password will only be given to the researchers of this project that need to access the data.

These data will be used only for analysis and for illustration in a PhD thesis, conference presentations and scientific journals. No other use will be made of them without your written permission, and no one outside the project will be allowed access to the original recordings.

What if I have any further questions?

If you have a concern about any aspect of this study, you should ask to speak to the experimenter who will do their best to answer your questions.

Children will be given the possibility to ask questions at the end of the study. Parents will also be given the researcher's contact details in case they have questions later.

If you would like more information about the study, please contact:

PhD student/Experimenter: Rui Huan

Email: r.huan.1@research.gla.ac.uk

Office: Room F132, Sir Alwyn Williams Building (School of Computing Science), Glasgow, G12 8RZ

Tel: +44 (0) 7746 8874 77

Supervisor: Professor Stephen Brewster

Email: Stephen.Brewster@glasgow.ac.uk

Office: Room S131, Sir Alwyn Williams Building (School of Computing Science), Glasgow, G12 8RZ

Tel: +44 (0) 1413 3049 66

Thank you for taking the time to read this information sheet.

Best wishes,

Rui Huan

Appendix E The smiley-o-meter questionnaire used for Chapter 4

Date of questionnaire:

Read this page before starting the questionnaire:

Hello:

You are invited to participate in our survey - estimating child engagement in animations. In this survey, you will be asked to complete a questionnaire that asks 8 questions about animation you just saw.

Each question includes a statement and 5 different smiling-face choice. You just need to choose one face based on what you really think. There are 4 different animations so you will be asked to complete 4 questionnaires.

Don't worry, we will show an example of one question to let you know how to fill this questionnaire.

You will need:

- a. a highlighter pen
- b. a pencil/pen
- c. a clipboard

Your participation in this study is completely voluntary. If you feel unhappy with answering any questions, you can withdraw from the survey at any point.

Thank you very much for your participation. Now Let's see an example of one question.

This is an example of one question:

Here is a picture. Please answer the question after this picture.








This is the statement. This question is asking your feeling to the picture above.

Q1. I liked this picture.

These are 5 smiling-face choice. You can choose one face based on your answer to the statement above.

Please use the highlighter pen or pencil to tick or fill one face.

				
No, I didn't like them at all.	No, I didn't like them.	I can't say either.	Yes, I like them.	Yes, I really like them!

Congratulations! You did really well!

Now you finished the example. Are you happy with starting our animation task now?

Story: Nightmare

Q1. I was absorbed in this animation.



No, I wasn't at all!



No, I wasn't really.



I can't say either.



Yes, I was quite.



Yes, I was totally!

Q2. I was involved in this animation that I'm happy to tell people what happens next.



No, I wasn't at all!



No, I wasn't really.



I can't say either.



Yes, I was quite.



Yes, I was totally!

Q3. I found this story confusing to understand.



Yes, it was really hard to understand.



Yes, it was a little hard for me.



It's OK.



No, it was easy to understand.



No, it was quite easy to understand!

Q4. When I was watching this story, I found myself thinking about other things.



Yes, I was totally thinking about other things!



Yes, I was quite thinking about other things.



I can't say either.



No, I wasn't really thinking about other things.



No, I wasn't thinking about other things at all!

Q5. I was stressed while watching this animation.



Yes, I was really stressed!



Yes, I was a little stressed.



It's OK.



No, I wasn't stressed.



No, I was stressed at all!

Q6. I felt I knew what the child doll were going through emotionally in this story.



No, I didn't
feel it at all!



No, I didn't
feel it.



I can't say
either.



Yes, I felt it.



Yes, I really
felt it!

Q7. I felt interested in this story task. (including the story and this questionnaire)



No, I didn't like
them at all.



No, I didn't like
them.



I can't say
either.



Yes, I like
them.



Yes, I really
like them!

Q8. How's the mummy doll feeling now? And what's the mummy doll thinking now?

Q9. How's the child doll feeling now? And what's the child doll thinking now?

Appendix F The smiley-o-meter questionnaire used for Chapter 5

Story: Nightmare

Q1. I was absorbed in this animation.



No, I wasn't at all!



No, I wasn't really.



I can't say either.



Yes, I was quite.



Yes, I was totally!

Q2. I was involved in this animation that I'm happy to tell people what happens next.



No, I wasn't at all!



No, I wasn't really.



I can't say either.



Yes, I was quite.



Yes, I was totally!

Q3. I found this story confusing to understand.



Yes, it was really hard to understand.



Yes, it was a little hard for me.



It's OK.



No, it was easy to understand.



No, it was quite easy to understand!

Q4. When I was watching this story, I found myself thinking about other things.



Yes, I was totally thinking about other things!



Yes, I was quite thinking about other things.



I can't say either.



No, I wasn't really thinking about other things.



No, I wasn't thinking about other things at all!

Q5. I was stressed while watching this animation.



Yes, I was really stressed!



Yes, I was a little stressed.



It's OK.



No, I wasn't stressed.



No, I was stressed at all!

Q6. I felt I knew what the child doll were going through emotionally in this story.



No, I didn't feel it at all!



No, I didn't feel it.



I can't say either.



Yes, I felt it.



Yes, I really felt it!

Q7. I liked the dolls and images (doll house and furniture) used on this story.



No, I didn't like them at all.



No, I didn't like them.



I can't say either.



Yes, I like them.



Yes, I really like them!

Q8. I liked the voice used on this story.



No, I didn't like them at all.



No, I didn't like them.



I can't say either.



Yes, I like them.



Yes, I really like them!

Q9. I felt interested in this story task. (including the story and this questionnaire)



No, I didn't like them at all.



No, I didn't like them.



I can't say either.



Yes, I like them.



Yes, I really like them!

Q10. How's the mummy doll feeling now? And what's the mummy doll thinking now?

Q11. How's the child doll feeling now? And what's the child doll thinking now?

Appendix G The instruction letter to labellers



The instruction letter of annotating the engagement level

Dear labellers,

Thank you for attending the brief MCAST training and you did really well!

This time you will be asked to rate the participant's engagement level during the following of the MCAST story-stems videos on a screen. According to the MCAST protocol (you already read during the brief MCAST training), the engagement level is a rating of representing the extent to which the participant has got absorbed and imaginatively caught up in the MCAST story-stems. It is rated by increasing attention to the story, lack of distraction to other things, and feeling empathy to the doll on the screen as seen by their facial expressions.

We need your judgement on the level of engagement of each video clip. The rating form was shown below:

Labeller:		
Allocated clips	Engagement level	examples of engaged/disengaged behaviours
1		
2		
3		
4		
5		

Your rating should be based on the description of engagement levels and relevant explanations that are provided to you. You need to give a single rating scale for each clip and write it down in the column named 'Engagement level'. Please be as objective and constructive as possible in your rating and use the following rating scale:

Level of engagement:

- 1 = not engaged, this clip shows that the participant was not engaged
- 2 = rarely engaged, this clip shows that the participant was rarely engaged
- 3 = highly engaged, this clip shows that the participant was highly engaged
- 4 = fully engaged, this clip shows that the participant was fully engaged
- X = If one clip was unclear (e.g. no eyes, eye/face occlusion) or contains no person at all

Also, we also need you to note examples of engaged/disengaged behaviours while realising during the observation. Please write them down in the last column named 'examples of engaged/disengaged behaviours'.

Please finish the annotation within 5 days and send the rating form back to me. Thank you very much!

Please feel free to contact me at the phone number/email if you have any questions about the annotation.

PhD student/Experimenter: Miss Rui Huan

Email: r.huan.1@research.gla.ac.uk Phone: +44 7746 8874 77