



**University of
Sunderland**

Hume, Colette (2017) Enhancing Questionnaire Design Through Participant Engagement to Improve the Outputs of Evaluation. Doctoral thesis, University of Sunderland.

Downloaded from: <http://sure.sunderland.ac.uk/id/eprint/7065/>

Usage guidelines

Please refer to the usage guidelines at <http://sure.sunderland.ac.uk/policies.html> or alternatively contact sure@sunderland.ac.uk.

**Enhancing Questionnaire Design Through Participant Engagement to
Improve the Outputs of Evaluation.**

Colette Hume

A thesis submitted in partial fulfillment of the requirements of the
University of Sunderland for the degree of Doctor of Philosophy

February 2017

ABSTRACT

Questionnaires are habitual choices for many user experience evaluators, providing a well-recognised and accepted, fast and cost effective method of collecting and analysing data. However, despite frequent and widespread use in evaluation, reliance on questionnaires can be problematic. Satisficing, acquiescence bias and straight lining are common response biases associated with questionnaires, typically resulting in suboptimal responses and provision of poor quality data. These problems can relate to a lack of engagement with evaluation tasks, yet there is a lack of previous research that has attempted to alleviate these limitations by making questionnaires more fun or enjoyable to enhance participant engagement.

This research seeks to address whether 'user evaluation questionnaires can be designed to be engaging to improve optimal responding. The aim of this research is to investigate if response quality can be improved through enhancing questionnaire design both to reduce common response biases and to maintain participant engagement. The evaluation context for this study was provided by MIXER, an interactive, narrative-based application for intercultural sensitivity learning, used and evaluated by 9-11 year old children in the classroom context.

A series of Participatory Design studies with children investigated engagement and optimal responding with questionnaires. These initial studies informed the design of a series of questionnaires created in the form of three workbooks that were used to evaluate MIXER with over 400 children.

A mixed methods approach was used to evaluate the questionnaires. Results demonstrate that by making questionnaire completion more enjoyable data quality is improved. Response biases are reduced, quantitative data are more complete and qualitative responses are more verbose and meaningful compared to standard questionnaires. Further, children reported that completing the questionnaires was a fun and enjoyable activity that they would wish to repeat in the future.

As a discipline in its own right, evaluation is under-investigated. Similarly user evaluation is not evaluated with a lack of papers considering this issue in this millennium. Thus, this research provides a significant contribution to the field of evaluation, highlighting that the outputs of user evaluation with questionnaires are improved when participant engagement informs questionnaire design. The result is a more positive evaluation experience for participants and in return a higher standard of data provision for evaluators and R&D teams.

Table of Contents

1	INTRODUCTION	11
1.1	<i>Research Question, Aims, Objectives and Rationale</i>	14
1.2	<i>Motivation</i>	15
1.3	<i>Structure</i>	16
1.4	<i>Summary</i>	18
2	LITERATURE REVIEW	19
2.1	<i>User Evaluation</i>	20
2.2	<i>User Evaluation Methods - Strengths and Limitations</i>	23
2.2.1	<i>Questionnaires</i>	24
2.2.2	<i>Interviews</i>	27
2.2.3	<i>Focus Groups</i>	31
2.2.4	<i>Observation</i>	35
2.2.5	<i>Biometric Methods</i>	37
2.2.6	<i>User-Centred Techniques</i>	39
2.2.7	<i>Questionnaires as the dominant method in user evaluation</i>	41
2.3	<i>Using Questionnaires in User Evaluations with Children</i>	43
2.4	<i>Understanding sub-optimal questionnaire responses</i>	46
2.5	<i>Engaging Users in Evaluation</i>	50
2.5.1	<i>Defining Engagement</i>	51
2.5.2	<i>Engagement Constructs</i>	51
2.6	<i>Involving Users in Designing Evaluation</i>	54
2.7	<i>Key Findings and Considerations from the Literature</i>	57
2.8	<i>Summary</i>	59
3	METHODOLOGY	60
3.1	<i>Positioning the Research</i>	61
3.2	<i>User Evaluation Questionnaires used in this research</i>	62
3.3	<i>Evaluating the User Evaluation Questionnaires - Mixed Methods and their application</i>	65
3.4	<i>Research Design</i>	66
3.4.1	<i>Phase 1: Review of Academic Literature & Children's Media: Inspiring Design</i>	76
3.4.2	<i>Phase 2: Preliminary Studies: Informing the Evaluation Design</i>	69
3.4.3	<i>Phase 3: Workbook Development: Implementing the Evaluation</i>	71
3.4.4	<i>Phase 4: Meta-Evaluation</i>	72

3.5	<i>Ethics, Recruitment and Consent</i>	77
3.6	<i>Summary</i>	77
4	PARTICIPATORY DESIGN STUDIES	78
4.1	<i>Basis and Scope of Studies</i>	79
4.2	<i>Questionnaire Design Workshop</i>	81
4.2.1	<i>Procedure</i>	83
4.2.2	<i>Questionnaire CDF - Results & Interpretation</i>	88
4.2.3	<i>Questionnaire CDF – Impact & Recommendations</i>	91
4.2.4	<i>Questionnaire Design – Results & Interpretation</i>	92
4.2.5	<i>Questionnaire Design - Impact & Recommendations</i>	96
4.3	<i>Five Degrees - Visual Likert Scale Development</i>	97
4.3.1	<i>Five Degrees Procedure – Study 1: Basic SFL</i>	99
4.3.2	<i>Five Degrees – Study 2: Visually appealing SFL</i>	102
4.3.3	<i>Five Degrees – Study 3: Neutral end point</i>	103
4.3.4	<i>Five Degrees – Study 4: exploring the negative end point</i>	103
4.3.5	<i>Five Degrees – Impact and Recommendations</i>	104
4.4	<i>Language Study</i>	105
4.4.1	<i>Procedure</i>	106
4.4.2	<i>Results and Interpretation</i>	107
4.4.3	<i>Impact and recommendations</i>	110
4.5	<i>Engaging Children with Questionnaires: Stickers as a response method</i>	111
4.5.1	<i>Procedure</i>	111
4.5.2	<i>Result and Interpretation</i>	112
4.5.3	<i>Impact and Recommendations</i>	113
4.6	<i>Visual questions & answers: Nine Square</i>	113
4.6.1	<i>Nine Square - Procedure</i>	114
4.6.2	<i>Nine Square – Results and Interpretation</i>	115
4.6.3	<i>Impact and Recommendation</i>	116
4.7	<i>Summary of findings</i>	116
4.8	<i>Summary</i>	119
5	INSTRUMENT DEVELOPMENT	121
5.1	<i>Instrument Development: Basic, Better & Best</i>	122
5.1.1	<i>Initial Instruments: Basic</i>	123
5.1.2	<i>Instrument assessment and refinement: “Better”</i>	123
5.1.3	<i>Transforming the Instruments: “Best”</i>	124

5.2	<i>Workbook Overview</i>	128
5.2.1	<i>Technical Development of the Workbooks</i>	131
5.3	<i>Workbook 1 – Pre-Test</i>	131
5.3.1	<i>Page 1 - Workbook 1 – Find the Camp</i>	132
5.3.2	<i>Page 2/3 - Workbook 1 - Messy: New Friendzzz</i>	133
5.3.3	<i>Page 4 - Workbook 1 - All packed and ready to go!</i>	134
5.3.4	<i>Page 5 - Workbook 1 – CQS – Which woodland animal?</i>	135
5.3.5	<i>Page 6/7 - Workbook 1 - Bryant’s Empathy</i>	136
5.3.6	<i>Page 10 - Workbook 1 - Summer Word Search</i>	138
5.4	<i>Workbook 2 - EEQ</i>	139
5.4.1	<i>Page 1 – Workbook 2 - Who wins?</i>	140
5.4.2	<i>Page 2 - Workbook 2 – Roving Reporter</i>	141
5.4.3	<i>Page 3 & 4 - Workbook 2 – True or False</i>	142
5.4.4	<i>Page 5 - Workbook 2 – MIXER</i>	143
5.4.5	<i>Page 6 – Workbook 2 - What do you think?</i>	144
5.4.6	<i>Page 7- Workbook 2 - iPad page</i>	145
5.4.7	<i>Page 8 - Workbook 2 - Friendship Word Search</i>	146
5.5	<i>Workbook 3 – Post test</i>	147
5.5.1	<i>Page 1 – Workbook 3 - New People, New Places</i>	148
5.5.2	<i>Page 2 – Workbook 3 - The Epic Quiz</i>	149
5.5.3	<i>Page 3 – Workbook 3 – Friends</i>	150
5.5.4	<i>Page 4 – Workbook 3 - Think Fast</i>	151
5.5.5	<i>Page 5 – Workbook3 - Maze Days</i>	152
5.5.6	<i>Page 6 – Workbook 3 – Spot the difference</i>	153
5.6	<i>Instrument Development & Transformation - Key Points</i>	153
5.7	<i>Summary</i>	155
6	RESULTS: Evaluating the Evaluation	156
6.1	<i>Measure One: Data Quality</i>	157
6.1.1	<i>Workbook Completion and Variance: Results and Interpretation</i>	158
6.1.2	<i>Workbook Completion and Variance: Key Findings</i>	160
6.2	<i>Measure Two: Engagement in Evaluation</i>	161
6.2.1	<i>Quantitative data collection via postcard</i>	162
6.2.2	<i>Qualitative data collection via postcard</i>	163
6.2.3	<i>Administering the postcard</i>	164
6.2.4	<i>Postcard Results: Quantitative Results and Interpretation</i>	164
6.2.5	<i>Postcard Results: Qualitative Results and Interpretation</i>	169

6.2.6	<i>Postcard Study: Key findings</i>	175
6.3	<i>Measure Three: CDF - Assessment of Engagement with Evaluation</i>	176
6.3.1	<i>Conducting the CDF</i>	177
6.3.2	<i>CDF: Results and interpretation</i>	178
6.3.3	<i>CDF: Key Findings</i>	179
6.4	<i>Summary</i>	180
7	DISCUSSION	181
7.1	<i>Synthesis of Research</i>	182
7.1.1	<i>Contribution from Literature</i>	183
7.1.2	<i>User Engagement in Evaluation Design</i>	189
7.1.3	<i>Discussion of Main Findings</i>	194
7.2	<i>Limitations & Considerations</i>	196
7.2.1	<i>eCute</i>	196
7.2.2	<i>Methodological Limitations and Considerations</i>	198
7.2.3	<i>Evaluator Role</i>	200
7.3	<i>Originality and Contribution to Knowledge</i>	202
7.4	<i>Future Work</i>	207
7.5	<i>Reflection</i>	210
7.6	<i>Summary</i>	211
8	CONCLUSIONS	213

List of figures

Figure 1.1: User needs supported by the R&D team	12
Figure 1.2: User supporting the R&D team during the evaluation phase	12
Figure 2.1: Tourangeau & Rasinski (1988), 4 stages of question answering.....	44
Figure 2.2: Smiley-o-meter (Read et al., 2002)	47
Figure 2.3: Thumbs-up scale (Kano et al. 2010)	47
Figure 3.1: Four-phase research design	67
Figure 3.2: Postcard - Quantitative data collection.....	75
Figure 3.3: Postcard - Qualitative data collection.....	75
Figure 4.1: An example section of a standard questionnaire document	82
Figure 4.4.2: Questionnaire Design Workshop Sessions and Tasks	83
Figure 4.3: Examples of questionnaire elements for children	86
Figure 4.4: Examples of the decorative designs created by the children	93
Figure 4.5: Creative use of the Likert scale format.....	94
Figure 4.6: Comic strip response item.....	94
Figure 4.7: Variety in response items used in questionnaires.....	95
Figure 4.8: The Non-Questionnaire	96
Figure 4.9: Early version of the PIL questionnaire	99
Figure 5.1: Three stage process to instrument development.....	123
Figure 5.2: Workbook 1 cover and participant data page.....	128
Figure 5.3: Workbook 2 cover	128
Figure 5.4: Workbook 3 cover	129
Figure 5.5: Find the camp	132
Figure 5.6: New Friendzzz	133
Figure 5.7: All packed and ready to go!.....	134
Figure 5.8: Which woodland animal are you?	135
Figure 5.9: Yes or No	136
Figure 5.10: The Trip.....	137
Figure 5.11: Summer word search	138
Figure 5.12: 'Who Wins' activity page and stickers	140

Figure 5.13: Roving Reporter	141
Figure 5.14: True or False activity and stickers.....	142
Figure 5.15: MIXER activity page	143
Figure 5.16: What do you think?	144
Figure 5.17: iPad page	145
Figure 5.18: Friendship Word Search	146
Figure 5.19: New people, new places	148
Figure 5.20: The Epic Quiz.....	149
Figure 5.21: Friends	150
Figure 5.22: Think Fast	151
Figure 5.23: Maze days.....	152
Figure 5.24: Spot the difference	153
Figure 6.1: Quantitative data collection to evaluate the workbooks	163
Figure 6.2: Qualitative data collected via the postcard.....	163
Figure 6.3: Histogram – Was the workbook fun to do (N=118)?	165
Figure 6.4: Histogram – Do you think the workbook looked good? (N=119).....	166
Figure 6.5: Histogram – Would you like another workbook to do in the future? (N=119).....	167
Figure 6.6: What the children didn't like about the workbooks – responses of children who didn't dislike anything are shown first in a darker shade	173
Figure 6.7: Reasons for disliking (with positive responses, i.e. there was nothing they disliked, shown in darker green).	175
Figure 7.7.1: Children applying stickers to the workbooks	192
Figure 7.2: Example pages from the final workbooks	194
Figure 7.3: Distinction between this research and the eCute Project.....	198
Figure 7.4: Model of evaluation informed by engagement	204

List of Tables

Table 2.1: Design and evaluation in the REVERIE project (Pasin et al. 2015)	21
Table 2.2: Data collection methods (Bargas-Avila & Hornbæk 2011)	23
Table 2.3: Summary of questionnaire methods.....	27
Table 2.4: Summary of interview methods	31
Table 2.5: Summary of Interview methods.....	34
Table 2.6: Summary of observational methods.....	37
Table 2.7: Summary of bio-metric methods	39
Table 2.8: Summary of user-centred methods.....	41
Table 3.1: Instruments used in the MIXER evaluation	64
Table 4.1: Summary of studies presented in this chapter	81
Table 4.4.2: Summary of CDF responses	91
Table 4.3: Summary of recommendations from questionnaire workshop	97
Table 4.4.4: Changes to Bryant’s Empathy Scale following the language study	107
Table 5.1: Far Transfer - Refining the Instruments	124
Table 5.2: Item of children’s media, purpose in research and where used in workbooks	127
Table 5.3: Evaluation activities in Workbook 1.....	130
Table 5.4: Activities in Workbook 3	130
Table 5.5: Overview of Workbook 2 (EEQ)	131
Table 6.1: Completion rates and variance in Workbook One.....	159
Table 6.2: Completion rates and variance in Workbook Two.....	159
Table 7.1: Preliminary Studies	189

1 INTRODUCTION

This research investigates if the outputs of evaluation with questionnaires are improved when participant engagement informs questionnaire design. This research takes a Participatory Design approach to designing engaging evaluation materials, by considering evaluation participants as end users of a product (questionnaires) and actively involving them in design and development.

In order to research, design, develop and test evaluation materials an evaluand is required. An evaluand is the subject of an evaluation, typically a program or system (rather than a person) (Scriven 1991). The evaluand used in this research, MIXER (Hall, Lufti, et al. 2011), is an interactive narrative prototype designed for children aged 9-11, providing a Serious Game that aims to support intercultural sensitivity learning. MIXER was developed within eCute (www.ecute.eu) an EU funded multidisciplinary Research & Development (R&D) project.

Evaluating an R&D prototype raises a particular set of challenges, in addition to the complexities faced with any standard product evaluation (Woodcock 2014). For example, the several disciplines typically involved in research projects will each have their own evaluation aims and requirements. When aggregated and applied during evaluation, participants face extensive and lengthy evaluation studies (Hall & Hume 2011). These evaluation studies are often intensive, requiring participants to complete several set tasks, interaction activities, discussions, and questionnaires to fulfil the evaluation

needs of the entire R&D team. R&D projects often apply Participatory Design (PD) or User Centred Design (UCD) methodologies, supporting the needs and requirements of the user by placing these at the heart of designing and developing interactive applications (Garrett 2010).

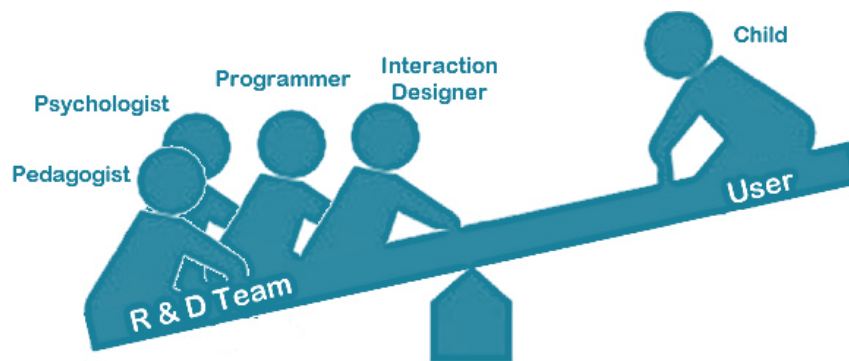


Figure 1.1: User needs supported by the R&D team

However, during evaluations the balance shifts and the focus is instead placed upon the needs and requirements of the R&D team. The user/evaluation participant, instead of being supported, becomes the support, bearing the weight of numerous evaluation requirements in the form of questionnaires, focus groups and interviews.

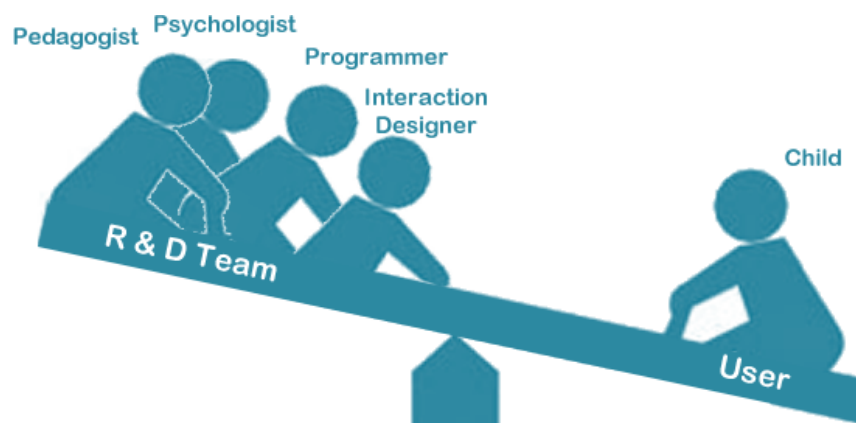


Figure 1.2: User supporting the R&D team during the evaluation phase

The problem with this model of evaluation for participants is that it can be uninteresting and tiresome, with more evaluation than interaction. The result is disengaged evaluation participants who seek to complete the evaluation task as quickly as possible, aiming to get back to what they were doing before, i.e. something more fun than evaluation...

Lack of engagement has a significant impact on data quality, with qualitative data not of the highest standard due to hurried and inaccurate interpreting and answering of questions (Krosnick et al. 1996). Where the users are children, as in the evaluation context of this research, this lack of engagement can result in children skipping questions that require more effort, such as free text/written elements (Zumbrunn et al. 2016). If children are disengaged then quantitative data may also suffer (Zaman et al. 2013). Child participants may satisfice (Krosnick 2000) and not provide true insights into their thoughts, feelings and opinions by succumbing to biases such as straight lining (Hall et al. 2016), acquiescence bias (Danner et al. 2015) or social desirability bias (Oerke & Bogner 2011).

Whilst engagement with interactive systems has been studied extensively (Linbo et al. 2015; Schoenau-Fog, 2011; Segel & Heer, 2010) it has not been considered in their evaluation. Through a series of Participatory Design studies this research aims to enhance evaluation materials, thereby improving user engagement with evaluation. As a result of improved engagement, it is hypothesised that occurrences of common questionnaire response biases will be reduced and thus data quality will be improved. Using a Participatory

Design approach for the design of evaluation materials is innovative, with a lack of previous research in this area.

1.1 Research Question, Aims, Objectives and Rationale

The hypothesis central to this research proposes that:

“The outputs of evaluation with questionnaires will be significantly improved when participants are engaged in the evaluation”

To engage participants in evaluation, this PhD seeks to answer the following research question:

“Are the outputs of evaluation with questionnaires improved when participant engagement informs questionnaire design?”

In exploring the hypothesis and research question, the aim of this research is:

To investigate if evaluation can be designed to engage evaluation participants and as a result gather high quality data by encouraging optimal responses and reducing occurrences of response bias.

To achieve this aim, the objectives of this research were:

1. To explore common response biases and children’s sub-optimal responding in questionnaire use.
2. To investigate constructs and measures of engagement and how they can be incorporated into questionnaire design to mitigate biases.
3. To involve users in evaluation design through applying participatory design approaches.
4. To create user evaluation questionnaires, incorporating key findings

from the literature on engagement and response biases and results from Participatory Design studies.

5. To administer the questionnaires and evaluate children's engagement with them, providing a meta-evaluation.

As a discipline in its own right, evaluation is much under researched and this PhD responds to notable gaps in user evaluation research. This research is amongst the first to consider participant engagement with evaluation and the impact of designing evaluations that engage participants. It is rare to see research that includes meta-evaluation (research that evaluates evaluation itself), with this thesis reporting a meta-evaluation, providing a valuable contribution to the little discussed subject of meta-evaluations. Finally, although questionnaires are frequently used in user evaluation, they receive little in the way of methodological consideration. This PhD addresses this, exploring the aesthetics of questionnaire design, the question comprehension and answering process and biases that can occur in questionnaire use.

1.2 Motivation

As a Research Assistant supporting the Sunderland team of the eCute project I was presented with many opportunities to follow in terms of research, these included innovative educational technology, novel interaction approaches, interactive narrative etc. But my interest was sparked by evaluation. I found evaluation to be a positive (seeking always to improve some aspect of life), useful, interesting, multifaceted and most significantly, a much under researched area. Contributing to the domain of evaluation, and casting light upon the repeatedly neglected intricacies and complexities of evaluation,

specifically relating to evaluation with questionnaires, was what inspired and motivated me to conduct this research in pursuit of the PhD.

1.3 Structure

The thesis is structured as follows:

Chapter 2 - Literature Review: The strengths and limitations of the main user evaluation methods (questionnaires, interviews, focus groups, observation, biometrics and user-centred methods) are reviewed. Optimal and sub optimal responses of children in questionnaires are considered examining causes of response bias and techniques to reduce these biases. Constructs and measures of engagement are reviewed, highlighting their potential for increasing participant engagement in user evaluation questionnaires. Participatory Design is explored as an approach to involving users in the design of engaging evaluation materials.

Chapter 3 - Methodology: This chapter outlines the methodological position underpinning this research and details the research context, outlining the reliable and valid user evaluation questionnaires used for the evaluation of MIXER. The four-phase research design applied in this research, using mixed methods and Participatory Design, is detailed exploring how the MIXER user evaluation questionnaires were designed and evaluated. This chapter also details the ethical approval, recruitment and consent processes relating to this research.

Chapter 4 - Evaluation development studies: Five small-scale studies that investigate engagement and response bias in evaluation are presented. The studies include a questionnaire focus group and design workshop (section 4.2), the incremental design and refinement of a 5 point Likert scale (section 4.3), a questionnaire language improvement study (section 4.4), an investigation of the use of stickers as a response format (section 4.5), and a study that aimed to improve the collection of qualitative data (section 4.6). These small-scale studies contributed to and informed the design of the final large-scale study. Method, results, recommendations and impact are provided for each study.

Chapter 5 - Instrument Development: This chapter details the design of 3 comic book style evaluation workbooks. The workbooks, which were designed to increase engagement and reduce response bias in evaluation, were used in the final large-scale evaluation of MIXER. The chapter details the design of each page/activity, indicating the engagement constructs and/or measures as well as any response bias that is addressed in that activity.

Chapter 6 - Meta Evaluation & Results: The chapter begins with an overview of the four measure approach applied in this research. The approaches combined qualitative and quantitative assessments of engagement through observation, a short feedback postcard, a classroom discussion forum, and assessment of data quality through

the data provided in the workbooks. The four measures are described and the results, interpretation and key findings from each measurement approach are provided. A summary of results concludes the section.

Chapter 7 - Discussion: This chapter provides a synthesis of the research presented in this thesis with a discussion of the various activities that contributed to the design, development and evaluation of the research presented. Limitations, originality, contributions to knowledge and future work relating to this research are discussed.

Chapter 8 - Conclusion: The final chapter of the thesis presents the main conclusions drawn from this research.

1.4 Summary

This chapter has introduced the background to and the focus of the research that will be conducted as part of this PhD. The hypothesis, aims, objectives and research question were provided, with the research question of the thesis being *“Are outputs of evaluation with questionnaires improved when participant engagement informs questionnaire design?”*

The motivation for this research and the contribution to knowledge was also included, highlighting the need for research in this area. Finally, a guide to the structure and content of the remaining chapters was provided. The next chapter provides a summary of the literature reviewed in this research.

2 LITERATURE REVIEW

This research focuses on improving user evaluation, with this chapter reviewing relevant literature and structuring this in the following sections:

2.1 User Evaluation: this section briefly discusses and defines user evaluation, highlighting that user evaluation occurs across the lifecycle, with a plethora of available methods.

2.2 User Evaluation Methods – Strengths and Limitations: this section focuses on user-oriented methods: questionnaires, interviews, focus groups, observation, biometrics and user-centred methods. It concludes that multiple methods are often used in user evaluations with the dominant user evaluation method being questionnaires.

2.3 Using Questionnaires in User Evaluations with Children: this section considers questionnaire use with children, detailing an optimal response model for question answering.

2.4 Understanding sub-optimal questionnaire responses: this section considers biases that can impact on children's question answering. It explores satisficing, acquiescence and social desirability, suggesting that children's questionnaire responses could be a result of the merger of these three biases.

2.5 Engaging Users in Evaluation: this section considers the constructs and characteristics of engagement that could be used to reduce the impact of satisficing resulting in optimal responses.

2.6 Involving Users in Designing Evaluation: this section reviews Participatory Design highlighting its potential as an approach to involving users in the design of engaging evaluation materials.

2.7 Key Findings and Consideration from the Literature: summarises the review and highlights the main inspirations from the literature.

2.1 User Evaluation

Evaluation is the formal, objective measurement and appraisal of the extent a given activity, project, or program has achieved its objectives (Zikmund et al. 2012). Scriven (2015) describes evaluation as the process of determining the merit, worth, or value of something. Whilst there is no standard definition of user evaluation, following Scriven, in this work, user evaluation is defined as “the process of determining the value of the user’s experience of an interactive (narrative) system.”

Evaluation is a critical phase in the user centred design process (Alkhafaji & Sriram (2012), in which the goals of the application are tested by collecting data for project partners and stakeholders to inform future development, report progress etc. However, this testing and collecting of data happens throughout the lifecycle informing design using a range of different evaluation methods and approaches, as highlighted in the following table from the REVERIE project (Pasin et al. 2015).

Phase	Technology Readiness Level	Design Goal	Technique	Involved Users
1	Narrative Scenario	Early user requirements collection	Online survey	Researchers, Online users
2	Prototype REVERIE Version 1 (RV1)	Formative Evaluation	Informal usability inspection, task analysis	Expert and potential users
3	Prototype RV2	Final version of user requirements	Cognitive walk through and lab test	Experts and potential users
4	Prototype RV3	Overall system pilot	Field trials	Real users

Table 2.1: Design and evaluation in the REVERIE project (Pasin et al. 2015).

With such a wide remit for evaluation, there are unsurprisingly a wide range of user evaluation methods: with 36 identified during a CHI workshop (Bevan 2009a). 100 detailed on the Autodesk blog (Autodesk 2016); 80+ user experience evaluation methods provided at (Rajeshkumar et al. 2013) (Roto et al. 2013); and many usability evaluation methods identified at (usability.gov 2013) (Bevan 2012). However, whilst apparently there may be many methods, these are primarily targeted at formative evaluation, used to inform and direct development, rather than to assess the summative value of the experience for the user.

Bernhaupt's (2015) categorisation of evaluation methods for games, identifies four evaluation method categories: user-oriented (e.g. questionnaires, focus groups, experiments, observation, etc.), expert-oriented (e.g. heuristic evaluation, expert walkthroughs), automated (e.g. telemetry analysis, accessibility tools) and specialised (e.g. atypical user-oriented methods for evaluating social engagement, coordination, etc.).

In this work, the focus is on user-oriented approaches, as whilst both automated and expert-oriented methods may provide useful data, it remains essential to evaluate with real users. For example, a recent comparison of web accessibility evaluation tools (Vigo et al. 2013) highlighted that automated testing alone is insufficient to evaluate. Similarly, with well-known expert-based approaches such as heuristic evaluation (Nielsen, J. & Molich 1990); cognitive walkthroughs (Wharton et al. 1994), group expert walkthroughs (Følstad 2007), best practice requires that expert-oriented evaluation methods are complemented through user evaluation.

Using user-oriented methods to assess and report the summative value of the user experience is widespread. The most frequently used approaches in the evaluation of interactive applications are self-report measures, in which subjects indicate subjective impression via rating scales or verbal reporting. In the majority of cases self-report methods are conducted with questionnaires, interviews, observation and focus groups. These methods can be used to evaluate before, during and after the interaction. Pre - post- designs are frequently seen in large-scale projects; however, smaller projects typically have one-off user evaluations, based on a single session.

User evaluation data is essential content for publications on interactive systems, with evaluation included in all presentations where an interactive experience is detailed in recent conferences (e.g. ACM's Interactive Digital Storytelling, Computer Human Interaction (CHI) and Interaction Design with Children). However, though there are many user evaluations reported, there

are relatively few different evaluation methods used, as detailed in the table 2.2 (Bargas-Avila & Hornbæk 2011).

Collection method	N	%*
Questionnaires	35	53
Interviews (semi-structured)	13	20
User observation (live)	11	17
Video recordings	11	17
Focus groups	10	15
Interviews (open)	10	15
Diaries	7	11
Probes	6	9
Collage or drawings	5	8
Photographs	5	8
Body movements	3	5
Psychophysiological measures	3	5
Other methods	18	27

*Notes. N=66 studies *data do not sum up to 100% because studies can use more than one method*

Table 2.2: Data collection methods (Bargas-Avila & Hornbæk 2011)

Since this review in 2011, user evaluation has seen an increased use of questionnaires with interviews, focus groups and observation typically used to substantiate quantitative findings. In addition, there has been increasing use of psychophysiological or biometric measures. In the following section the main user evaluation methods are briefly discussed, highlighting strengths and limitations.

2.2 User Evaluation Methods - Strengths and Limitations

User-oriented methods involving the user in summative user testing are essential for user evaluation and the focus of this thesis, with its central

question of “*Are the outputs of evaluation with questionnaires improved when participant engagement informs questionnaire design?*” The main summative user-oriented evaluation methods (interviews, questionnaires, focus groups, observation, user-centred (Nacke 2015), each with strengths and weaknesses, are briefly detailed in the following sections.

2.2.1 Questionnaires

A questionnaire is a self-report measure usually with written questions which aim to extract specific information from the chosen respondents (Kaplan 2015). They are the most widespread and well-known approach for user evaluation and are well understood as a means of gaining subjective opinion by most people.

Questionnaires offer the flexibility to utilize a variety of response formats and data types collected (Harlacher 2016). A questionnaire may use scales only collecting quantitative data, another may only contain open-ended questions gathering qualitative data and another may combine both, enabling some triangulation through mixed methods.

Structured closed-question formats are commonly seen in user evaluation. These include: Dichotomous scales, which provide opposing statements such as ‘Yes-No’, ‘Agree-Disagree’, ‘True-False’ as response statements (Birkett 2015). Semantic Differential scales, using polar adjectives such as ‘hot-cold’, ‘strong-weak’, ‘happy-sad’ which are arranged at either end of a continuum. Respondents are asked to indicate where on the continuum their agreement

with the given statement lies (Sanoff 2016). Likert scales are another example of a commonly used structured response item that has been used for many year to gather attitudes or opinions (Likert 1932) and is extensively used in user evaluations.

Questionnaires used in user evaluation include the use of well validated and widely used measures for collecting psychological data such as the Ten Item Personality Index (TIPI), (Gosling et al. 2003), Bryant's Empathy Index (Bryant 1982b) and Hofmann et al. (2016)'s Interpersonal Emotion Regulation Questionnaire (IERQ). Questionnaires have also been developed for directly assessing aspects of the user experience, for example AttrakDiff (Hassenzahl 2004) and the Aesthetic Scale Lavie & Tractinsky (2004). Specific questionnaires have been proposed to measure the experience of people interacting with computer games (Klimmt, Roth, Vermeulen, Vorderer, & Roth, 2012), or interactive narrative (Yannakakis et al. 2013) focusing on factors, such as engagement, enjoyment, flow, playability and many others. In addition, there are many self-developed user evaluation questionnaires although there are concerns over the reliability and validity of results generated from their use (Bevan 2009b).

Wolff et al. (2016) states that an advantage of structured formats such as closed-question questionnaires is a lower cognitive load on the respondent, which leads to higher response rate and more accurate data. Additionally coding and analysis is faster and easier when a structured format is used (Timpany 2011). However, there is also often a need for open-ended

questions that allow for a more nuanced insight into respondents opinions than is possible when using a scale of some sort (Spool 2015).

Open-ended question are intended to encourage richer, more detailed answers using the subject's own knowledge and/or feelings (Harlacher 2016). However, respondents often dislike providing written feedback and will provide minimal content (Zumbrunn et al. 2016) or miss the question out completely. A further challenge for the open question format is that during coding in the analysis phase, there is opportunity for subjectivity by the researcher and as such results may be prone to bias (Bryman & Bell 2015)

Questionnaires can provide the option for anonymity of respondents where necessary (Melián-González 2016). This may encourage respondents to reply more openly and honestly, improving the quality of the data collected, particularly in sensitive subject areas such as exclusion or sexual and aggressive behaviours as explored in some Serious Games.

Questionnaires do have long recognised weaknesses, for example, Ackroyd (1992) argued that questionnaires are inadequate to understand some forms of information, i.e. changes of emotions, behaviour, feelings etc. The use of scales and pre-defined categories can limit users' ability to express themselves. Questionnaire design is critical with Coombe & Davidson (2015) noting that fatigue and a lack of engagement can occur if a questionnaire is too long or monotonous. Such lack of engagement can result in users losing motivation and providing random rather than considered responses.

Whilst there are some weaknesses, the many advantages of questionnaires, primarily that they are a well-recognised and accepted measurement approach that enables large scale data collection quickly, with low effort and cost have resulted in their extensive use in user evaluation. Questionnaires are a well established method viewed as the most appropriate by many researchers and practitioners and reported in many publications.

Questionnaire Strengths	Questionnaire Limitations
Practical and effective, with low cost and effort.	Inadequate to understand some forms of information i.e. changes of emotions, behaviour, feelings etc.
Highly flexible with the possibility utilize a variety of question types and response formats,	Questionnaire design must be high quality and of user appropriate length.
Many validated and reliable questionnaires already exist	Lack of motivation to complete questionnaire can result in incomplete or poor quality data.
Provide the option for anonymity of respondents where necessary.	
Structured formats can lower cognitive load on respondents - leads to higher response rate, more accurate data and that coding and analysis is faster and easier	

Table 2.3: Summary of questionnaire methods

2.2.2 Interviews

An interview is a qualitative data elicitation method, where an interviewer asks questions and the interviewee responds either verbally or by text, depending on the type on interview. Interviews have been used extensively across many disciplines (Gubrium & Holstein 2012) for centuries, with their purpose to gain a deeper understanding of interviewees' perspectives (Wyse 2014).

In user evaluation, interviews are typically used to gain users' self-reported and subjective views of the interaction they have just experienced (Raita

2012). Most interviews happen directly after interaction, providing a well-known user evaluation method (Wilson 2014). User evaluation interviews are usually synchronous in time and can be differentiated by the level of structure (e.g. structured, semi-structured and unstructured) and the medium through which the interview occurs (e.g. face-to-face, phone, Instant message/chat, in a virtual space), with semi-structured face-to-face interviews the most typically seen in user evaluation. User evaluation interviews are typically recorded, supplemented with interviewer notes, then transcribed and analysed, often using content and thematic analysis.

The main strengths of face-to-face interviews for user evaluation is in the generation of rich, considered, qualitative data relating to the user's experience. This rich data is obtained through eliciting respondent's views and experiences in their own terminology rather than limiting their options to a pre-defined set of choices (Kaplan & Maxwell 2005). Whilst interviews do generate spontaneous responses, with the focus on discussing subjective views, interviews can also encourage more reflection by the user on their interactive experience through appropriate probing questions (Sutcliffe & Hart 2013). In addition to rich verbal data, interviews also capture non-verbal communication, with Wyse (2014) noting that non-verbal cues such as body language can indicate discomfort with the questions asked, or conversely, enthusiasm for a subject being discussed.

User evaluation interviews provide a systematic, tailored and flexible approach to exploring the user's experience. During an interview, the

evaluator has the potential to clarify questions for the respondent (Schober 2016), to explore unanticipated responses and views and to probe for additional information (Opdenakker 2006).

Whilst interviews have many strengths, a key weakness can relate to the interviewer. Interviewer effects such as the evaluator guiding the interviewee in the direction they wish them to go or poorly designed interviews can flaw the interview. Training, use of interview protocols and ensuring evaluator awareness of potential effects and how to mitigate them, can reduce interviewer effects. Interviewers in semi-structured interviews (as typical of user evaluation) need to develop particular skills, including “double attention” (Wengraf 2003) that is the listening to the user’s responses to understand the perspectives being provided and formulating subsequent questions both to maximize the input of the user and to cover all the evaluation areas in the available time.

There are also interview weaknesses related to respondents for example with interviewees particularly prone to the response bias of social desirability (Dahlgren & Hansen 2015). Or, participants who might feel uneasy about the anonymity of their responses, thus tailoring them. Users’ opinions can change during interviews as they reflect on their experience (Sutcliffe & Hart 2013). This can be positive, however, it can also be the result of biases such as social desirability or acquiescence, with participants providing the views that they think the evaluator wants to hear. However, if the interviewer has been appropriately trained, then they can establish rapport and an effective

ambience, underpinned by a considered interview protocol then such issues are reduced.

Interviewer and interviewee aside, the main weaknesses or challenges of interviews is the time, effort and cost needed to undertake them. With a large number of respondents face to face interviews can prove costly when time taken and travel costs are considered (Marshall 2016). In aiming to reduce this cost dimension, remote evaluation has been used, with a number of tools enabling user evaluation interviews, from phones to Skype.

Whilst video conferencing / phone-based approaches do have some of the strengths of co-located interviewing, there is some reduction in richness, with Seitz (2016) noting that the use of Skype can result in an *“inability to read body language and nonverbal cues, and loss of intimacy compared to traditional in-person interviews”*. Castro & Gramzow (2015) compared face-to-face and webcam interviews, with their findings highlighting the possibility that researchers conducting webcam interviews may misjudge non-verbal cues from respondents.

Irvine et al. (2012) compared telephone interviews to face-to-face interviews, finding that telephone interviewees were less confident that the information they were providing was meeting the researcher's needs, and, unsure of the response required, were less forthcoming in their responses. However, Shapka et al.'s (2016) comparison of data quantity and quality in online interviews versus in person interviews with adolescents found that while the

online chat interviews produced fewer words and took longer to complete, data quality was unaffected by the mode of data collection.

A potential challenge for interview data relates to analysis and the time and effort required for this. However, similar time and effort is required for any data analysis. Like other forms of data analysis, qualitative analysis is supported by tools, such as speech recognition software, speeding transcription and Nvivo, reducing analysis time. It does remain challenging to analyse and interpret the analysis of such rich, qualitative data, however, this in itself is not an inherent weakness, rather something that has to be factored into planning. The weakness may be that it is not.

Interview Strengths	Interview Limitations
Rich verbal and non-verbal data giving deeper understanding of users' perspectives	Face to face interviews can be expensive in terms of financial costs, time and effort.
Flexible and tailored to the user, using their terms and enabling them to direct focus of interview	Respondents limited by cost and time restrictions.
Potential to clarify questions for the user and responses for the interviewer	Interviewer effect and interview ambience
If necessary, can be undertaken remotely with only some reduction in data richness.	Distributed interviews can impact on richness of data

Table 2.4: Summary of interview methods

2.2.3 Focus Groups

A focus group is a group interview technique that benefits from communication between participants in order to generate qualitative data (Stewart & Shamdasani 2014). Focus groups have many of the benefits of interviews, with Kitzinger & Barbour (1999) stating that focus groups are *“Particularly useful for allowing participants to generate their own questions,*

frames and concepts and to pursue their own priorities on their own terms, in their own vocabulary". In addition, Silverman (2016) observes that when respondents hear the input of others in the group this often triggers thoughts and ideas that wouldn't have occurred otherwise. Thus, Focus groups have the particular advantage in user evaluation of stimulating reflection of the user's experience.

The suggested ideal number of participants in a focus groups varies from 6-12 (Nielsen 1997; Trochim & Donnelly 2006; Freeman 2006). While these numbers vary slightly, it is agreed that smaller groups allow for an easy flow of communication between group members while remaining controllable by the facilitator.

The main advantage of Focus groups in comparison to interviews is that they can provide rich qualitative data whilst requiring little in resources (in terms of time, manpower and cost) (Krueger & Casey 2014). Focus groups are used widely in product development and evaluation (Fuller 2016; Dickinson et al. 2016),

The success of the focus group relies upon the skills of the facilitator to keep participants focused and to ensure that everyone has their say. Niyonzima (2015) comments that focus groups can be intimidating at times, especially for inarticulate or shy members. Overbearing group members are also a problem to be aware of when conducting focus groups. An overly dominant group member could make other members feel less confident about contributing, and a particularly enthusiastic or vocal participant may sway the opinions of

others in the group (Traynor 2015). Focus groups (similar to interviews) can also be affected by social desirability bias, for example participants may answer in a certain way to appear more appealing to researchers etc., (Oerke & Bogner 2011).

A skilled facilitator will promote and control debate and at times challenge participants in order to draw out peoples true thought, feelings or opinions (Krueger & Casey 2014). As stated a benefit of focus groups comes from the interplay of discussions between participants, however, there is an assumption here that participants will be quite verbose in their exchanges with each other and the facilitator. Where this is not the case the success of a focus group largely relies on the skill and experience of the facilitator, (Carey & Asbury 2016), who will ensure the aims of the session are achieved.

Getrich et al. (2016) state that researchers often fail to describe in detail the complexity of conducting focus groups, including what ensues when the unexpected occurs. And indeed, where Focus Groups are reported in the literature, they are, as Getrich et al. note, often portrayed as:

“...a controlled scientific endeavour, in which the perfectly constituted sample is chosen, an exact number of participants actually show up as planned, the ideal set of questions are crafted ahead of time and deployed with precision by the moderators, the conversation among similarly situated participants flows naturally and smoothly, and the data generated from the encounter are ultimately the end-result of the exercise.”

There are few publications that discuss focus group implementation, with most focusing exclusively on results. This, however, occludes one of the major advantages of focus groups, that is their flexibility and adaptability to context, user group and experience. For example, Hall et al. (2004) in the evaluation of the 'Fearnot!' application by 9-11 year olds developed Classroom Discussion Forums. These provide a classroom-centric focus group that met teacher preference to support the whole class and small group activity typical of the classroom. The CDF was tailored to the age group with many researcher short questions with raised hands for answering, rather than a more general discussion. The authors conclude that tailoring a focus group to context had high ecological validity and generated invaluable input from a child-centred perspective for the design of FearNot.

Focus Group Strengths	Focus Group Limitations
Hearing the input of others often triggers thoughts and ideas that wouldn't have occurred otherwise	Some members can dominate whilst others are too shy to speak – this can be mitigated by effective facilitation.
Allows participants to generate their own questions, frames and concepts and to pursue their own priorities on their own terms, in their own vocabulary	Requires skilled facilitator to keep the session focused and to ensure the aim of the session is achieved
Requires little in resources (in terms of time, manpower and cost) compared to other techniques	Confidentiality of participants contributions cannot be guaranteed once the session has ended
Can be adapted to a wide range of contexts, having considerable flexibility.	

Table 2.5: Summary of Interview methods

2.2.4 Observation

In observation methods the researcher watches the evaluation participants during an interaction. Observations can be structured or unstructured. In a structured approach the researcher uses a list of criteria or benchmarks, these can be ticked off or assigned with a predefined score as the study progresses (Bryman & Bell 2015). A benefit of a structured approach is that quantitative data, as in other methods, is faster and easier to analyse. In an unstructured approach the researcher simply makes notes on their observation of the participants interaction.

Observation evaluation methods are flexible and applicable across a wide range of user experiences. Observation can be conducted in field or lab based studies, depending upon the focus of the study (Wilson & Sharples 2015). Khanum & Trivedi (2012) list three commonly applied approaches of observation: Direct Observation, Think Aloud and Constructive interaction.

In Direct Observation the evaluator can take notes, use pre-defined templates and/or record the interaction for later coding and analysis. Direct observation typically involves co-location of evaluator and participant, although video recording is increasingly used. In addition to video, logging provides a form of direct observation, with technologies enabling the synchronisation of multiple observational streams.

Ferreira et al. (2016) describe the Think Aloud method as follows, '*Users are asked to literally think out loud and report their interaction, the tasks they are*

performing and what difficulties they are having". Thus, it is possible to obtain data about the users' reasoning during the performed tasks. User evaluation beyond a commentary on actions has used Think-aloud style methods to gauge users' engagement and enjoyment, for example in the measurement of Continuation Desire (Schoenau-Fog 2011).

Constructive interaction is where two participants work in a group while observed by the researcher. Nielsen (1994) states that constructive interaction is more effective over think-aloud when conducting usability evaluations with specific user groups, such as children. Where children face difficulties in following the instructions for a think-aloud test, constructive interaction comes closer to their natural behaviour, allowing the children work in pairs and collaborate in solving the tasks.

Observation does have a number of potential weaknesses, including the Hawthorn effect, also known as observer effect. McCambridge et al. (2014) describe observer effect as participants' awareness of being studied, and a possible impact on behaviour. This can be mitigated to some extent by ensuring that participants understand the purpose and focus of the evaluation. Similar to interviews, observations, with their focus on individual user sessions, can be costly and time consuming both to administer and analyse.

A key limitation of direct observation, both logged and recorded, is that alone this approach does not give insight into the user's decision process or attitude during the interaction (Merriam & Tisdell 2015). Where this is supported as in methods such as Think Aloud, these are intrusive and interrupt the user's

experience, with Khanum & Trivedi (2012) noting that during think aloud users may not feel able to fully carry out the assigned task while simultaneously communicating a commentary of their actions. Further, as noted by Ganglbauer et al. (2009) that whilst eliciting information about participants emotional state (when measuring enjoyment for example) is crucial, asking users about their emotional state often means interrupting the flow of the interaction and experience.

It is unusual for observation to be used alone for evaluation and it is typically used in conjunction with other evaluation methods (Portell et al. (2015). For example, Merriam & Tisdell (2015) state that observation is best used for triangulation with interviews or questionnaires to substantiate findings.

Observation Strengths	Observation Limitations
Enables evaluator to observe users and their interaction, potentially providing use	Direct observation offers little insight into users thoughts and decision making processes
Structured observation can gather quantitative data which is easier to analyze	Think Aloud participants may struggle to carry out the task and provide a commentary
Useful when used with other methods to substantiate findings	Reliability is an issue in observation – two observers are recommended
Can be used in field or lab studies	Participants may be subject to observer effect

Table 2.6: Summary of observational methods

2.2.5 Biometric Methods

Biometric methods are based on the measurement of physiological responses of users during an interaction. The use of biometrics is increasing in user evaluation, particularly in video games and VR applications (Bian et al. 2016; Wiemeyer et al. 2016; Wu et al. 2015). A range of physiological evaluation

approaches have recently become more widely available including eye tracking (Kruger et al. 2016) gaze duration (Georges et al. 2016), galvanic skin response (Mundell et al. 2016), heart rate (Bian et al. 2016) and pressure exertion on input device (Quax et al. 2013) along with approaches to support data interpretation (e.g. Fera-2015 (Valstar et al. 2015), Facial Action Coding System (Craig et al. 2008)

Like direct observation, a key advantage of physiological measures is that interruption is not necessary when physiological methods are employed, as data can be collected continuously during the interaction (Rawassizadeh et al. 2015). However, as with observation, again the challenge remains of understanding what that data might mean and how it should be used to evaluate the user experience. Whilst biometric measures do require specialist equipment, this has reduced dramatically in cost in recent years. The increased use of self-monitoring equipment for health and fitness has changed user's perception of physiological measures, with equipment and environment in which physiological studies are conducted no longer 'unnatural for the participant' (Dirican & Göktürk 2011). There are technical challenges with collecting biometric measurements, with some measures easily affected by external influences, e.g. pupil dilation may be effected by light in the environment rather than a psychophysiological reaction (Chen et al. 2016).

Although challenging, biometric testing has become increasingly popular, with a significant increase in user evaluations incorporating some biometric testing. However, with the challenges of analysing the data, as with observation, it is

very unusual for user evaluations to consist only of physiological measures and typically additional measures (e.g. interviews, questionnaires) are used to complement physiological data.

Biometric Strengths	Biometric Limitations
Biometric measures are more difficult for users to control deliberately so may help avoid the act of social masking	Some measures are easily affected by external factors. For example, pupil dilation
Increasingly accepted by users with growth in physiological self-measurement devices	Biometric measures can need specialist equipment which is often expensive and requires training and such lab setting can have an effect on participants
	Data can be difficult to interpret and may not provide insights into why users responded as they did.

Table 2.7: Summary of biometric methods

2.2.6 User-Centred Techniques

There are a range of user evaluation methods influenced by user-centred design approaches, tailored and designed for specific user groups and contexts. They aim to provide alternative approaches to gathering and generating data typically using techniques that are more interactive and engaging than questionnaires, interviews and focus groups.

Examples include eMoto, (Fagerberg et al. 2004) an emotional text-messaging service, Affective Diary (Lindström et al. 2006), which provides a mirror of a user's bodily experiences during the day by capturing sensor data in combination with text and MMS messages and photographs, or AffectCam (Sas et al. 2013), a wearable system capturing the user's galvanic skin response along with photos collected during the day. Probes, (Colombo & Landoni 2014) photography (Behrendt & Machtmes 2016), body-storming

(Murchú 2016) and drawings (Ferreira et al. 2016) are also seen. Transmedia Evaluation (Hall & Hume, 2011; Hall et al., 2013, 2014) is a user-centred evaluation methodology that applies user-centred techniques including participatory design to the design of the entire evaluation, creating in-game evaluation through notes, drawings, questionnaires, interviews and focus groups, experienced by the user as part of the interaction.

The main advantage of user-centred approaches is that evaluation materials are designed to elicit the user experience in a way that meets the user's expectations. This approach engages users in reflecting about their experience. Using diaries, video diaries, drawings, taking photos, etc. is typically viewed positively by users with work on probes highlighting that participants enjoy receiving and using probe elements.

Evaluation methods using user-centred design approaches can be effective. However, they are more difficult to design and administer than more traditional methods, such as questionnaires and interviews. They can be time-consuming and the data can be difficult to aggregate and represent as an outcome of the evaluation.

As with observations and biometric measures, non-standard techniques are typically complemented by additional evaluation methods, most frequently questionnaires. Results from non-standard techniques are typically used to triangulate, substantiate and provide further evidence for quantitative results.

User-centred Methods - Strengths	User-centred methods Limitations
Tailored to the user matching their expectations.	High effort and time in developing innovative evaluation methods.
Users highly motivated to engage with the approaches, with many having a sense of fun	Can be difficult to represent results in formats appropriate for publication
Can provide rich data providing reflective perspectives.	Can be difficult to generalize from results as method strongly tailored to context and users.

Table 2.8: Summary of user-centred methods

2.2.7 Questionnaires as the dominant method in user evaluation

As has been detailed in earlier sections there are strengths and weaknesses in all user evaluation methods. To mitigate these weaknesses and maximise strengths, in many user evaluations more than one evaluation method is used. This use of multiple methods for user evaluation is acknowledged as being likely to improve evaluation results (Sutcliffe & Hart 2013).

Many user evaluations report an observed interaction followed by relatively short questionnaires complemented by an additional qualitative approaches, typically interviews or focus groups. This can be seen in evaluations of a forensic Serious Game (Drakou & Lanitis 2016) and an interactive digital storytelling experience in a museum (Rizvic et al. 2012) etc. There is also increased use of bio-metric measurement as an additional method, for example with galvanic skin response, pressure exertion, questionnaire and “informal oral discussions” being used to investigate games accessed via networks (Quax et al. 2013).

Where multiple methods are used, almost all user evaluations include questionnaires. Further where single methods are used for evaluation, these

are most likely to be questionnaires. Questionnaires dominate user evaluation (Barkhuus & Rode 2007), with Vermeeren et al.'s, (2010) review of evaluation methods in 92 studies identifying that questionnaires were the dominant choice and used in over two thirds of the studies reviewed. Inargas-Avila & Hornbæk's (2011) review of 51 empirical research papers, questionnaires were identified as being the prevailing evaluation method with 35 of 51 studies using quantitative questionnaires. This trend appears to be on the increase, with over two thirds of studies reviewed in Ólafsson, Livingstone, & Haddon's review of studies of children's use of the internet only collecting quantitative data and few studies using mixed methods.

Questionnaires are viewed as a scientific, rigorous and valid method to gather data. There is general consensus that questionnaires are an effective method for evaluating users, applying formats and approaches adopted from disciplines such as psychology. However, as Bevan comments "For nearly a century, survey researchers have, for the most part, designed questionnaires in an intuition-driven, ad hoc fashion," and this is regularly seen in user evaluation with many self-developed questionnaires with dubious validity (Bevan 2009b). Although questionnaires are becoming ubiquitous in user evaluations, they have received little methodological consideration although there are concerns about how useful questionnaires and their results are in user evaluation (Robertson 2012).

In the domain of interactive media evaluation (and in research generally) there are a lack of papers and practitioner experiences about how evaluations are

designed and iterated (Hall et al. 2016). Often, authors will state that the questionnaire/study was piloted but fail to give any detail on the process followed. Van Teijlingen & Hundley (1998) note there is need for more discussion among researchers of the process and outcomes of pilot studies.

Thus, whilst almost every interactive narrative publication includes an evaluation involving a questionnaire, there has been almost no consideration of how questionnaires should be designed for use as user evaluation instruments. This continues the trend noted by Raita (2012): *“in the 21st century papers that concentrate on the analysis and development of evaluation methodology have almost disappeared.”*

In this thesis, the focus is on this most used of the user evaluation methods: questionnaires. There is considerable heterogeneity in user evaluation questionnaires depending on the purpose, users, context and application being evaluated. In this thesis, the focus is on using questionnaires to capture data from children aged 9-11, with the following section focusing on the challenges of gaining optimal responses using questionnaires with children for user evaluation.

2.3 Using Questionnaires in User Evaluations with Children

A key issue for questionnaires is that the accuracy of the data depends upon the users' performance of a series of cognitive processes in answering the questions (Vannette & Krosnick 2014). To achieve an optimal response

requires consideration of questionnaire answering in general rather than user evaluation. Here, the issue is not how the user judges the interface (e.g. as would be modelled following Hartmann's approach (Hartmann & Sutcliffe 2008) for example) but rather how this perception can be captured using questionnaires.

Gaining an optimal response in questionnaire answering is frequently represented as a four stage model: question comprehension, information retrieval, summary judgment and response communication (e.g. see Cannell, Miller, & Oksenberg, 1981; Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000). In this thesis, the approach taken is that provided in Tourangeau & Rasinski, (2000)'s seminal evaluation work: *The Psychology of Survey Response*.

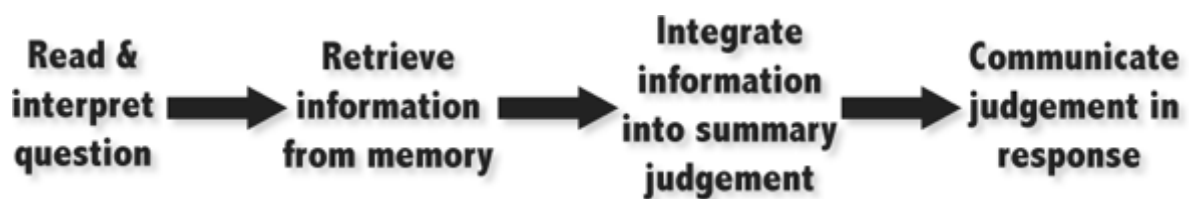


Figure 2.1: Tourangeau & Rasinski (1988), 4 stages of question answering

Bell, (2007) notes that for a child to provide an optimal response the following must be true:

1. The child must be able to understand the words and the sentence that forms the question statement
2. The child must be able to associate the question statement with a past experience of their own in order to retrieve the required information to complete step 3
3. The child must understand that the questionnaire is asking them to

make a judgment of their past experience against the question statement

4. The child must be able/provided with an effective method to communicate the judgment made in step 3

There has been extensive work demonstrating the need to create understandable, usable questionnaires, which children can read, interpret and respond to. For example Larsen et al., (2008), state that particular consideration must be given to the audience for which the questionnaire is intended to ensure its language and content is appropriate. Factors that impact on question answering include developmental effects including language ability, reading age, and motor skills, as well as temperamental effects such as confidence, self-belief and the desire to please, (Read & MacFarlane, 2006). Whilst there are still poorly designed questionnaires administered to children, this is an easily solved issue with growing awareness of the need to provide age-appropriate questions and aesthetics.

In user evaluation, the second of the stages, information retrieval is facilitated by the child having taken part in an interaction. This interaction provides the child with a past experience of their own to enable them to answer the questions in the user evaluation. Where questionnaires are used that focus on attitudes and opinions (e.g. psychological measures) if the question is understandable and tailored to the age group, then children should be able to make optimal responses.

Children understand the concept of questionnaires, thus comprehend that they need to answer the question using the response format provided,

meeting stage 3. Assuming an appropriate scale is used that is understandable for the child, it should then be possible for them to communicate their judgement. Studies have shown that in communicating judgements to questions that children prefer Likert scales over similar simple response items such as Visual Analogue Scales and there has been considerable use of Likert scales in evaluating with children (Mellor & Moore 2014; Haddad et al. 2012)

Although this would suggest that an optimal response should be achieved, there is growing awareness that children's responses are not always optimal (Zaman, Vanden Abeele, & De Grooff, 2013). The following section further considers why sub-optimal answering may be occurring when questionnaires are used for user evaluation with children.

2.4 Understanding sub-optimal questionnaire responses

Increasingly children's user evaluation questionnaires are well designed with appropriate language, scales and aesthetic. They are often piloted and refined, tailored to the age group etc. However, many child evaluation studies demonstrate extreme positive results (der Sluis et al. 2015) with child respondents agreeing or strongly agreeing to scaled questions, with some children straight-lining (Cole et al. 2012), that is, ticking all the boxes down one side of the page of a questionnaire. Throughout the literature such positive results are interpreted as showing that the interactive system is

engaging, easy to use, entertaining, etc. Whilst there is some reflection on such positive results, few researchers flip the issue and question whether the children's judgements were optimal.

However, sub-optimal responding has long been recognised as a problem with attempts to address this issue. For example, the work of Read et al. (2002) in creating the Smiley-o-meter (figure 2.2) or the Thumbs-Up Scale by Kano et al. (2010) (see figure 2.3) were both attempts to increase the quality of responses to questions from children.



Figure 2.2: Smiley-o-meter (Read et al., 2002)



Figure 2.3: Thumbs-up scale (Kano et al. 2010)

However, even using such child-centric scales, there are doubts about whether children are providing an optimal response due to inconsistencies between questionnaire results and qualitative findings. For example Zaman, Vanden Abeele, & De Grooff, (2013), found that the Smiley-o-meter (Read et al., 2002) produced results that were inconsistent with children's actual product preferences. Additionally, Mellor & Moore's, (2014), more recent study on the use of Likert scales with children concluded that children have a

limited understanding of the use of Likert response formats. And as noted their use can result in children straight lining and extreme responding (der Sluis et al. 2015), almost always with highly positive outcomes.

Whilst the results of child evaluations with questionnaire are positive this does not mean that they are optimal. For example, Bell, (2007), explains children tend to use the easiest route possible to create an answer that they feel satisfactorily meets the requirements of the task, the less effort the better, meaning question quality becomes even more important with this group.

This 'less effort the better' approach to questionnaires is referred to as satisficing. Satisficing is a cognitive bias in which respondents decide on and carry out (either consciously or unconsciously) a course of action that will satisfy the minimum requirements necessary to achieve a particular goal and is the opposite to an optimal response (Krosnick 1991; Vannette & Krosnick 2014). Jäckle & Eckman (2016) describe satisficing as "*respondents pick an easy credible answer, instead of processing the question optimally and answering truthfully.*"

Krosnick et al. (2015) states that satisficing may take many forms for example, selecting the first reasonable response to avoid reading the rest of the provided options, agreeing with assertions or a lack of differentiation in ratings where scales are provided (e.g. straight lining and/or extreme responses), Vannette & Krosnick (2014) further explain that the extent to which satisficing takes place is likely to be related to and dependent on three related key factors, as detailed in Krosnick's formula.

$$P (\text{Satisficing}) = \frac{A_1 (\text{Task Difficulty})}{A_2 (\text{Ability}) \times A_3 (\text{Motivation})}$$

Equation 2.1: Krosnick's (1991), formula of satisficing

Krosnick states that of the three factors listed above, although the respondent's ability is out of the researchers control, the remaining two (task difficulty and motivation) can be influenced to reduce the sensitivity to satisficing. Krosnick provides recommendations for reducing satisficing including maximising respondent motivation (e.g. describe the purpose and value of the study; provide instructions to think carefully); minimising task difficulty (e.g. minimise the number of words in questions and maximise the familiarity of words used); and minimising response effects (e.g. offer responses in balanced or random order; avoid agree/disagree, true/false, yes/no questions).

In addition to the satisficing bias, there are various additional biases that could have an impact on enabling children to make optimal responses. For example social desirability bias may result in participants not accurately responding to questions regarding socially desirable characteristics in order to appear more appealing to researchers etc. (Oerke & Bogner 2011). Acquiescence Bias or the a tendency of respondent's to agree or respond positively to questions if in doubt (Danner et al. 2015) is also seen. In user evaluations with children, the two biases can merge, with children wanting to appear more socially desirable through positively responding to the system being evaluated. The tendency of

children to demonstrate extreme positive results can be attributed to some degree to the interaction between satisficing (least effort required) and acquiescence / social desirability (desire to agree with and please adults).

Whilst it is possible to reduce social desirability and acquiescence through an appropriate protocol, satisficing, or responses made with the least effort possible, provides a significant issue for user evaluation questionnaires with children. A potential approach to reducing satisficing and increasing optimal responses is to increase the child's motivation by improving their level of engagement with the questionnaire. However, there has been little consideration of how user evaluation questionnaires can be designed to engage children, (or indeed adults), nor of the impact of engagement on optimal responses. In the following section, approaches to engagement are considered, aiming to identify potential ways of improving optimal responding.

2.5 Engaging Users in Evaluation

In the design and development of interactive, narrative based, learning applications great amounts of time and effort have been spent to achieve novel and exciting experiences that engage and enthuse users. Recent research into engagement has covered a range of domains, from E-commerce (O'Brien 2010), faculty community involvement (Wade & Demb 2009), museum exhibits (Black 2005; Tjøstheim et al. 2015) to computer

games (Schoenau-Fog 2012), investigating and developing methods to make these experiences more and more valuable and engaging for users.

2.5.1 Defining Engagement

O'Brian & MacLean, (2009), define engagement as a quality of user experience that facilitates more enriching interactions with computer applications with the researchers focusing on two main issues related to engagement: fun and finance. Engagement related to finance focuses on how people can be engaged to spend more, stay longer, buy more, come back and buy again etc. and is not relevant to the focus of this research. However, engagement related to fun is of considerable relevance with applications targeting children almost always intending to be enjoyable and to generate a positive experience.

2.5.2 Engagement Constructs

Engagement is a product of many constructs (Attfield, Piwowarski, Kazai, et al. 2011), for example, an interaction could be engaging because it invokes one or more of the following constructs e.g. satisfaction (Yannakakis 2008; Aguirre et al. 2014), enjoyment (Weber et al. 2009; Tjøstheim et al. 2015), fun (Bartle 2004; Tasci & Yong 2015), usefulness (D'Mello et al. 2012), meaningfulness (Schoenau-Fog, 2011), pleasure (Douglas & Hargadon 2000; Chiewvanichakorn et al. 2015), novelty (O'Brien & MacLean 2009), immersion (Cairns et al. 2014) or motivation (D'Mello et al. 2012; Wang et al. 2008).

As well as individual constructs there are also models of engagement, including process models, such as Sharafi, Hedman, & Montgomery (2006) who specify five modes of engagement: Enjoyment / Acceptance; Ambition / Curiosity; Avoidance / Hesitation; Frustration / Anxiety and Efficiency / Productivity. Other models define the characteristics of engagement and in this research, the model used is that of Attfield, Kazai, Lalmas, & Piwowarski, (2011) who identify eight characteristics in their model of engagement: focused attention; positive affect; aesthetics; durability; novelty; richness and control; reputation, trust and expectation; and user context. Of these, five are particularly relevant to designing children's evaluation questionnaires:

- **Novelty** appeals to our sense of curiosity, encourages inquisitive behaviour and promotes repeated engagement (O'Brien & Toms 2010). In evaluation this sense of novelty encouraging repeat engagement is particularly useful. For example, when the use of a pre and post use questionnaire is required children may become uninterested in answering the same questions and the likelihood of satisficing, straight lining, or acquiescence bias may increase. Providing novel questionnaire activities and aesthetics could increase optimal responding.
- **Positive affect / fun** is a key element of engagement. Features of 'Fun' are described as challenge, curiosity and fantasy in game play (Law 2011). Read, MacFarlane, & Casey's, (2002), toolkit for measuring fun with children described three dimensions of fun, Endurability, Engagement and Expectation. Providing these features in evaluation materials could have a significant impact on children's engagement.
- **Aesthetic Appeal.** Much of the literature reporting aesthetic appeal relates to interface design in contexts including online shopping, web

search, educational content and video games (O'Brien & Toms 2010). However, the aesthetics in interface design relates to factors such as layout, graphics used and the application of design principles such as symmetry, balance and use of colour, (Attfield, Piwowarski & Kazai 2011; Short et al. 2015) are equally applicable to questionnaires. Another factor of the aesthetic design of the evaluation materials is the layout of the questions, with the potential to remove linearity from the layouts, potentially reducing straight lining.

- **Endurability / Returnance.** Attfield, Piwowarski, et al., (2011) describe endurability as remembering an experience and being willing to repeat it. Read, MacFarlane, & Casey, (2002b) use the term returnance to describe a desire to repeat an activity that has been fun. Returnance will be the term used to refer to the desire to repeat an experience throughout this research. Returnance will be incorporated into the evaluation of the user evaluation questionnaire, as a measure of how many children would want to repeat the evaluation experience.
- **Focussed attention** assesses if the application holds the users attention, (Matlin 1994; Qiong 2015). In an evaluation context do the evaluation materials focus the participants and hold their attention long enough to ensure that all elements of the evaluation are complete? Can a well-designed and engaging questionnaire gather a 100% complete data set? Completion rates will be used as a measure of focused attention as a construct of engagement.

Creating engaging experiences requires involvement of the user within the design process. It is standard for researchers, designers and developers to work closely with users and stakeholders by actively involving them in the design of interactive applications, applying methods including participatory design, co-design and co-creation. It would seem likely that to create

engaging questionnaires the respondents also need to participate in the questionnaire design.

2.6 Involving Users in Designing Evaluation

Participatory design, increasingly referred to as co-creation, is a user-centred approach in which stakeholders, end-users, designers and researchers contribute to the design process in order to help ensure that the end product meets the needs of its intended user base (Anić 2015). Participatory design is a well-established approach used in technological product development (Simonsen & Robertson 2012).

Participatory design distinguishes itself from other approaches by acknowledging and involving users as experts with knowledge of the context that the technology will become part of (Bratteteig & Wagner, 2014). Participatory design strives to broaden the perspective of and increase empathy in design by giving specific and often under represented user groups, a voice in the design process (Chisik & Mancini 2016).

Druin et al. (1998), who pioneered participatory design with children, identified a spectrum of roles: users, testers, informants and design partners. Most focus has been on the latter, with participatory design approaches considering users as *partners* or co-designers in the design process rather than merely informants (Robertson & Wagner, 2012). However, whilst design-partnering has been viewed as the best way to gain user input (DeSmet et als. (2016)

review of Serious Games in Health identified that the informant role was more effective than input from the user throughout the design process.

Participatory design is seen as an effective way of improving the user experience and in recent years has greatly diversified with a broad spectrum of approaches and methodologies emerging (Frauenberger et al. 2015). However, although Participatory Design is widely used, formal evaluations of Participatory Designs are rare with a lack of details on the methods used (Bossen et al. 2016). Consequently, there is little evidence as to whether or not Participatory Design makes the experience effective.

Although in Portnoy et al.'s (2008) review of 75 studies of Serious Games for health were shown to be more effective when participatory design was used, the opposite finding was seen in DeSmet et al.'s (2016) meta-analysis of 61 health-related Serious Games studies. This meta-analysis highlighted that Participatory Design was associated with higher effectiveness when it was applied to game dynamics, levels, and game challenge rather than when it was applied to game aesthetics.

Although there is a lack of clear evidence that participatory design improves game effectiveness, there is agreement that Participatory Design does lead to better user experiences. Frauenberger et al. (2015) claim that more relevant and meaningful technology can be created by giving people who are affected by it a role in its design. Other benefits of Participatory Design include reducing the risk of failure as designs are based on real facts and findings provided by users themselves rather than assumptions.

As with other collaborative methods, designers need to facilitate rather than prescribe during the Participatory Design process. There can be challenges with Simonsen & Robertson (2012) highlighting problems in finding and recruiting suitable participants and in acquiring their on-going (possibly long term) commitment to the project. And, as with any collaborative approach, conflicts can occur between participants with successful participatory design relying upon creating good relationships between the designers and users/co-designers.

As stated by Anić (2015) a participatory design approach aims to ensure that the end product meets the needs of its intended end user base. Although Participatory Design is used extensively where the end product is an application or piece of technology developed to address the needs of the user in completing a task, it has had little use in the design of user evaluations. A user evaluation questionnaire is an end product; yet, unlike the products it is being used to evaluate it is not user-centred nor designed with the user's involvement.

Although Kusunoki & Sarcevic (2012) propose the use of Participatory Design for designing evaluation methods, the focus is on how Participatory Design can be used to identify *what* needs to be evaluated, rather than on using Participatory Design to design a user evaluation. For example, with the questionnaires the authors intend to develop, they state "*we will gather expert feedback on the technical development of the questions and instruments. This will help ensure that survey design recommendations are followed.*" However,

it isn't clear what input (if any) the users have to the design of the questionnaire.

With the exception of eCute's Transmedia Evaluation (Hall et al. 2015; Hall & Hume 2011) which I contributed to, there is almost no work on using Participatory Design to create evaluations designed with, and for, the user. Participatory Design has clear potential for designing user evaluation instruments, such as questionnaires. The research presented in this thesis explores how involving children in the design of evaluation questionnaires impacts upon optimal responding.

2.7 Key Findings and Considerations from the Literature

Most interactive narratives are evaluated using user evaluation methods including questionnaires, interviews, focus groups, observation, biometrics and user-centred approaches. Evaluations often use multiple user evaluation methods to mitigate weaknesses and to gain additional insights. However, questionnaires are the dominant user evaluation method, used in the majority of studies with adults and children, and are the focus of this research.

The tendency for positive results in interactive narrative studies highlights that evaluations with children are often achieving little differentiation between participants when data is collected by questionnaire. Such sub-optimal responses are a significant issue in children's evaluation questionnaires with serious implications for the quality of the collected data and the validity of the

results. This research explores how optimal responding can be improved for child evaluation questionnaires.

Most child evaluation questionnaires are well designed, comprehensible and age appropriate. However, the considerable number of studies reporting only positive results suggests that the challenge in gaining optimal responses lies deeper than simply in the use of language and an age-appropriate aesthetic, with biases resulting in sub-optimal responses. Key approaches from the literature review that will inspire and impact upon the research approach and design to improving optimal responses in children's evaluation questionnaires by increasing engagement are:

- 1. Optimal Response Model** – this research will use Tourangeau & Rasinski's (2000) four stages of question answering to explore how optimal responding can be achieved with children.
- 2. Satisficing Formula** - Krosnick's (1991, 2000) work on satisficing was also influential to this research. Krosnick lists three key factors to consider in reducing satisficing 1) task difficulty, 2) respondent ability, 3) respondent motivation. These will be addressed by reducing task difficulty, creating materials that match participant ability and finally by creating evaluation materials that engage and as a result motivate.
- 3. Engagement** - Attfield, Kazai, Lalmas, & Piwowarski, (2011)'s model of engagement with its characteristics of focused attention; positive affect; aesthetics; durability; and novelty; will be applied in the design of the evaluation materials, aiming to create an engaging evaluation that will motivate children to provide optimal responses.
- 4. Participatory Design** – Participatory Design offers considerable potential for improving user engagement with evaluation materials, such as questionnaires. An informant approach will be used aiming to

increase user engagement with the questionnaire and achieve optimal responses through user-informed design.

2.8 Summary

This chapter has reviewed the main areas of focus in this research. It has reviewed user evaluation methods, identifying strengths and weaknesses across a range of user-oriented evaluation methods. This review identified that many user evaluations employ multiple evaluation methods, however, that the dominant method is questionnaires. The use of questionnaires with children (the user group for the research in this thesis) raised concerns about data quality, with biases such as satisficing, social desirability and acquiescence having an impact on children's responses. The potential of engagement to reduce response biases was highlighted, with a review presented of engagement characteristics that could improve motivation and optimal responding. Finally, Participatory Design was identified as an approach that could contribute to improving engagement in questionnaires through involving participants in their design. The following chapter details the methodology used to explore how children's engagement with questionnaires could be improved to increase optimal responses.

3 METHODOLOGY

This chapter presents an overview of the methodology used in this research, presenting a mixed methods approach, providing quantitative and qualitative methods enhanced and refined through applying Participatory Design techniques and approaches. This chapter is presented as follows:

3.1 Positioning the Research: Briefly outlines the philosophical position (positivist, empirical, hypothesis based, transdisciplinary) that underpins the methodological approach for this research.

3.2 User Evaluation Questionnaires: this section briefly discusses the valid and reliable user evaluation questionnaires that were used in the user evaluation of MIXER (the evaluand).

3.3 Mixed Methods and their application: this section discusses the selection and suitability of the mixed methods approach.

3.4 Research Design: describes the four-phase research design that was applied to the design, implementation and evaluation of the MIXER evaluation, detailing how data quality and engagement of the user with the evaluation was assessed.

3.5 Ethics, Recruitment and Consent: Provides ethical considerations, recruitment and consent procedures followed in this research.

3.1 Positioning the Research

This research takes a positivist approach; with empirical studies exploring, investigating and assessing user engagement with evaluation. A positivist approach has been defined as being empirical, scientific, objective and hypothesis based (Creswell 2012). The hypotheses and questions resulting from this positivist approach provide the empirical framework upon which to scaffold the studies and exploration of engagement and evaluation.

A key factor in positioning this research is the user-centred, participatory design approach taken throughout. The empirical studies have been designed to occur in a real world context (e.g. a classroom or school). As detailed below this has had a significant impact on the selection of methods and the design of materials, with consideration being given to the user as well as to the needs of the R&D team for whom the evaluation is providing results.

This research is transdisciplinary, Scriven (2015) describes a transdiscipline as a discipline that focuses on issues essential to other disciplines but has in itself the attributes of a discipline. This transdisciplinary perspective infuses the approach to the use and selection of methods requiring mixed methods, with input from a range of disciplines (e.g. computing, psychology, education, media). Further it highlights the potential for the findings and outcomes of this research to have relevance across many disciplines and evaluation contexts.

3.2 User Evaluation Questionnaires used in this research

This thesis focuses on the use of questionnaires as a user evaluation method, investigating how optimal responding can be improved. In user evaluation, as detailed in section 2.2.7, questionnaires are the dominant method. Validated questionnaires can provide a valid and reliable method of accurately measuring that which is to be measured (Saunders et al. 2009). However, in user evaluation, questionnaires are often self-developed to fit a specific 'one off' purpose. In their review of evaluation studies Vermeeren et al., (2010) note that a number of questionnaires they reviewed were of "*questionable scientific quality*" because of a lack of validation studies, whilst (Zaman et al. 2012) have highlighted doubts about the validity of many evaluation results due to the concerns about the provenance of the questionnaire.

There are many books that provide approaches to designing questionnaires (e.g. Oppenheim's (1992) or Tourangeau & Rasinski's (2000) seminal work), with agreement that iterative development with piloting or pretesting is an essential step in development. Piloting is essential to ensure that questions are understandable, ensuring that data analysis techniques match expected responses and to evaluate the reliability and validity of the instrument (Kitchenham & Pfleeger 2002).

The validity of a questionnaire relates to how well it measures the intended research concept or construct that it is meant to measure. Reliability is also critical, determining whether the measurement is consistent and stable, and thus, allows others to repeat the measure. Developing valid, reliable

questionnaires is challenging and involves significant testing and refinement. Even so, there are many well-validated and reliable measures, that when used in user evaluation can be assumed to be equally reliable and valid.

In the context of this research, the focus was not on creating the constructs, questions and scales used to evaluate the users' learning and experience of a Serious Game. Instead, the focus was on how the user evaluation method could be implemented to increase optimal responding. The four questionnaires used to evaluate MIXER, a Serious Game for learning intercultural sensitivity, were specified by the eCute project team for use in a pre- in- post- test design, see figure 3.2.

The pre- and post- test measures used 3 reliable, validated questionnaires to measure the effectiveness of MIXER in learning intercultural sensitivity, as detailed in the following table.

Learning Goal	Questionnaire	Rationale
EMOTIONAL Be able to recognise emotions (e.g. fear and anxiety) when dealing with the novel / unknown behaviours of another group?	Cultural Intelligence Scale (CQS) (Ang et al., 2007)	The behavioural subscale of the CQS is used as a pre and post measure of a child's capability to adapt verbal and nonverbal behaviour in different situation/cultures. This will provide data for the question: "Do children who have a more flexible repertoire of behavioural responses in culturally diverse settings recognise more emotion/behaviours in the MIXER application?" This will address aspects of the behavioural and emotional learning outcomes.
COGNITIVE Start learning the specific practices and values of that group?	Bryant's Empathy Index (Bryant, 1982)	Factor One from the Bryant Empathy Index will be used as a measure of children's empathic behaviour and styles. This will provide data for the question: "Are children with higher empathy levels more able to recognise and accept emotions in novel situations?" This will address the emotional goal of the learning outcomes: "Be able to recognise your emotions when dealing with strange behaviours of another group".
BEHAVIOURAL Being fully present in attending to others verbal and non-verbal messages.	MESSY (Matson Evaluation of Social Skills) (Matson et al., 2010)	Factor two and four of the MESSY questionnaire have been selected to determine children's capability to adapt to verbal and nonverbal behaviour in different situations/cultures to assess the behavioural goal from the learning outcomes: "Be fully present in attending to others verbal and nonverbal messages".

Table 3.1: Instruments used in the MIXER evaluation

The fourth questionnaire, the Experience Evaluation Questionnaire (EEQ) has been used in various formats since 2004 (Hall, Woods, Dautenhahn, et al. 2004) to evaluate Serious Games. Although this questionnaire has not undergone the rigorous development of the other three questionnaires, it has been used with 2500+ children and adults in evaluating a wide range of interactive narratives, including virtual agents (Hall et al. 2005), synthetic characters (Endrass, Hall, Hume, Tazzyman, Andre, et al. 2014) and social robots (Hastie et al. 2016). For MIXER, the EEQ (Hall et al. 2014) included questions that evaluated the Agents and Experience, Engagement and Interaction and Learning and Comprehension, these are further described in the following table 3.3.2.

Evaluation Focus	Evaluation
Agents and Experience	Agent believability and effectiveness, including presentation, communication and mind architecture.
Engagement and Interaction	The level of engagement experienced by the user in respect to their interaction with MIXER and the characters. Usability, user experience and enjoyment evaluation of the interaction approach
Learning and Comprehension	The level of the user's understanding of the events and progression of the scenario and interaction. Emotional, cognitive and behavioural learning

Table 3.3.2: EEQ Evaluation Goals

The questionnaires used for the MIXER user evaluation were reliable and valid. They appropriately measure a range of constructs including empathy, social skills, cultural intelligence, comprehension and experience. Using validated, reliable questionnaires allows this research to focus on the evaluation method itself, rather than the questions that it asks, the typical focus of questionnaire development; or the responses that it generates, the typical focus of user evaluation.

3.3 Evaluating the User Evaluation Questionnaires - Mixed Methods and their application

Mixed methods aim to answer research questions by making use of appropriate previous research and/or more than one type of investigative method by combining qualitative and quantitative data collection techniques (Haq 2015; Zikmund et al. 2012) The use of mixed methods enables a

detailed exploration of engagement, gaining data and experience from users in a range of studies.

In addition to enabling the use of a variety of approaches collecting a range of data offers particular benefits to this research as they can provide a deeper understanding of interactions by allowing the triangulation of data (Webb et al. 1966). Triangulation is often cited as having methodological superiority over single methods (Symonds & Gorard 2008), increasing validity when multiple findings either confirm or confound each other (thus reducing the chances of inappropriate generalisations) (Haq 2015). As detailed in chapter 2, mixed methods are frequently seen in user evaluation; for example, self report techniques, interviews, questionnaires, attitude scales, interaction logging, personality inventories and observation.

A mixed methods approach was used to evaluate the user evaluation questionnaires that were created in this research. This triangulated quantitative questionnaire data with qualitative data obtained through a questionnaire, observation and a focus group; these are further detailed in the Research Design below.

3.4 Research Design

In this research, Participatory Design and extensive piloting were used to explore how optimal responding could be increased, looking at designing out biases such as satisficing (see section 2.4) and increasing user engagement,

followed by the application of mixed methods to assess the questionnaires. All of the participatory design and empirical work with children occurred *in the wild*. This aimed to maintain ecological validity of the situation for the children and avoid engaging in interactions and evaluations in a non-real situation. Also, to address social desirability bias, the setting of the evaluations was to be as natural as possible for the children. Thus, all studies detailed in this thesis were conducted in the classroom.

The purpose of this PhD was to investigate if the outputs of evaluation with questionnaires are improved when participant engagement informs questionnaire design. To achieve this, a four-phase research design was followed:

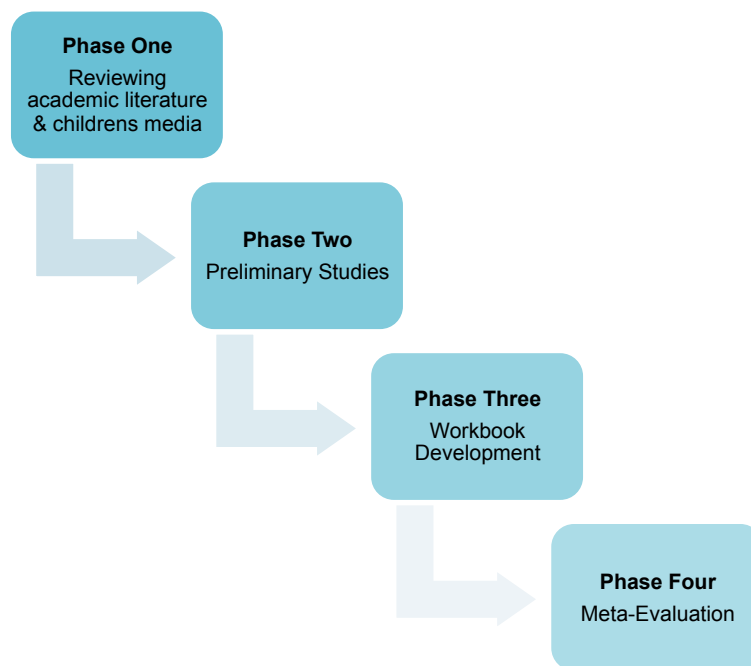


Figure 3.1: Four-phase research design

3.4.1 Phase 1: Reviewing Academic Literature and Children's Media / Inspiring Design

In addition to consulting academic literature (see chapter 2) to gain an understanding of the theory behind evaluation, response bias and engagement, a review of media targeted to the age group of 9-11 year olds was also conducted. The aim of this review was to inform the design of the Participatory Design activities and the questionnaires by understanding the design practice applied in the production of children's media.

The reviewed media included both educational and recreational literature, as these are the two most frequently accessed forms of literature by 9 to 11 year olds. Educational media included various SATs and Key stage 2 support materials in both print and online. However, as fun and enjoyment were crucial design features in terms of engagement, recreational media was the main focus of the review. There was also a lot more variety in the recreational literature which included comic books / magazines, special interest magazines (e.g. Doctor Who, Bird Life etc.), activity books / sheets, websites, sticker collecting books, fiction and non-fiction books and annuals.

The review involved reading and interacting with hard copy media targeting children focused on content, activities, aesthetics and integration with narrative. The design inspirations from the review are provided in Appendix H.

3.4.2 Phase 2: Preliminary Studies: Informing Evaluation Design

Phase 2 aimed to investigate:

- What aspects of evaluation *are* and *are not* engaging to children?
- How can high quality data be collected in a way that is engaging for children?

To investigate this, Phase 2 involved the use of a range of Participatory Design methods aiming to stimulate and engage children in designing evaluation with questionnaires, including:

- **Participatory Design Workshops & Early Stage Design Techniques** - Participatory Design workshops and activities are extensively used in interactive narrative system design (Bødker et al. 2016; Frølund 2014) with findings and results often having a significant impact on future interaction design (Hazelden 2007). Participatory Design workshops were selected for use in this research, with the aim of generating ideas and outputs that could be used to create more engaging questionnaire designs. Participatory Design methods used were based on those used for early stage design with the aim of creating low fidelity prototypes (e.g. creating storyboards (Rice, Cheong, Ng, Chua, & Theng, 2012), engaging in role play (Wang et al. 2015), scenario development (Johnson et al. 2012). Participatory Design was also used to explore constructs of engagement and their impact on evaluation, with a focus on what

engaged the children and what they particularly enjoyed.

- **Gaining Children's Qualitative Opinions** - In addition to engaging in design, children's qualitative opinions were obtained using (Hall, Woods, Dautenhahn, & Sobreperéz, 2004)'s Classroom Discussion Forums (CDF), a classroom-centric focus group approach. The CDFs were used to explore children's views of, and ideas about, questionnaires and to investigate children's opinions of the evaluation materials produced.

This initial exploration, further detailed in chapter 4, resulted in the development of questionnaire prototypes. These were then used in a series of studies investigating engagement with the aim of reducing and eliminating the biases identified in chapter 2 (e.g. satisficing, social desirability and acquiescence). The approach taken to investigate the questionnaires replicated that typically used to evaluate interactive narrative systems, however, in this research the system is replaced by the questionnaire:

- Introduction to questionnaires and purpose of evaluation (e.g. to evaluate and improve the questionnaire)
- Interaction with the questionnaire and / or individual elements.
- Evaluation of the questionnaire / elements using quantitative approaches including questionnaires and qualitative elements including user-generated drawings, texts and storyboards.
- Classroom Discussion Forum about the questionnaire / elements

The approaches and results from Phase 2 are further detailed in Chapter 4, with some of the studies that explored, trialled and evaluated the questionnaires generating results in publications on: developing the MIXER storyline (Hall et al., 2011) evaluating interaction modalities (Endrass, Hall, Hume, Tazzyman, Andre, et al. 2014); understanding children's views of conflict resolution (Hall et al., 2012). These studies provided a useful vehicle to further inform the final design of the questionnaires as detailed in Phase 3.

3.4.3 Phase 3: Workbook Development / Implementation

In Phase 3, the questionnaires were designed based on the findings from the Participatory Design of Phase 2. The questions used in the questionnaires are detailed in section 3.2.

In Phase 3, the questionnaires were re-designed, based on the results of the Participatory Design aiming to increase engagement and as a result 'design out' sub-optimal responses, through incorporating design elements that had reduced biases and/or increased engagement in the preliminary studies of phase 2. This 3-stage re-design of the questionnaires from basic (questionnaires in their original basic form) to better (question sets reduced, language improved) to best (transformation to increase engagement and reduce response bias) was an iterative process. The final questionnaires provided as workbooks (see chapter 5) underwent a constant process of iterative refinement using Participatory Design approaches with children. Phase 3 ended with the deployment of the questionnaires in the summative evaluation of MIXER for the eCute project:



Figure 3.2: MIXER Summative Evaluation

Although the questionnaires used for the user evaluation were critical to this thesis, the actual results and what they evidenced are not. In this thesis, the results that are reported relate to the user's evaluation experience, rather than the data collected in the evaluation. The results and data from studies using the questionnaires can be found in eCute publications and deliverables (see www.ecute.eu).

3.4.4 Phase 4: Meta-Evaluation

In Phase 4, the questionnaires were evaluated, providing a meta-evaluation conducted from 2 perspectives: the researcher's perspective – that is the quality of the data generated by the questionnaire; and the user's perspective – that is children's engagement with the questionnaire.

The meta-evaluation involved the following steps:

- Children completed the user evaluation questionnaires in their classroom
- Observation of the children whilst they completed the user evaluation questionnaires.
- After user evaluation questionnaire completion the children completed a post-card questionnaire gathering quantitative and qualitative data (see figures 3.2 and 3.3)
- Finally, children participated in a Classroom Discussion Forum generating qualitative data.

The quality of the data gathered by the four user evaluation questionnaires was assessed in terms of

Completeness: that is did the children answer the questions and complete the questionnaires. It was hypothesised that if children were not engaged by the questions then they would be less likely to answer them with fewer questions completed if children were not engaged.






Individual Variance: that is did each child use the entire range of scale points across the questionnaires. This indicated whether children were optimally responding or just straight-lining. Higher individual variance would suggest higher engagement and that the child had thought about the question providing an optimal response.

Sample Variance: this explored whether the sample as a whole used the entire scale. It was hypothesised that if children were engaged and providing optimal responses, sample variance would be higher.






Engagement was assessed through a mixed methods approach with a postcard questionnaire (see figures 3.2 and 3.3). A postcard questionnaire was selected as the format for the short data collection exercise that followed workbook completion. The postcard format was selected as it continued the trip/holiday theme of the MIXER application that was applied throughout the evaluation materials and activities.

A postcard was selected, as they are a recognised method of collecting feedback data in a variety of situations and locations, for example, restaurants, health care, libraries and theme parks, and thus should be familiar to children. The postcard was designed as a short data collection exercise, giving consideration to the fact that while the workbooks were designed to be fun for the children it was important that the children should not be over burdened with too many tasks. The two-sided postcard consisted of eight questions, three quantitative and eight qualitative. This questionnaire evaluating the user evaluation questionnaire was deliberately short as children had already answered a significant number of questions in the user evaluation of MIXER. Qualitative data was collected with the questionnaire as well as through observation and a Classroom Discussion Forum.

HOW DID WE DO?
 Please tell us what you think about todays session.
 Your name.....
 Was the workbook fun to do?

				
Very Much	Quite a lot	It was OK	Not really	Not at all

Do you think the workbook looked good?

				
Very Much	Quite a lot	It was OK	Not really	Not at all

Would you like another workbook to do in the future?






				
Very Much	Quite a lot	It was OK	Not really	Not at all

Figure 3.2: Postcard - Quantitative data collection

What was your favourite activity in the workbook?

Why did you like it the best?

What didn't you like about the workbooks?

Why didn't you like it?

What would make the workbooks better?

Figure 3.3: Postcard - Qualitative data collection

The quantitative questions focused on children's opinions of the fun and the appearance of the workbooks and their desire to further engage with the workbooks. This quantitative data was analysed using descriptive statistics

and Spearman's Rho (correlation coefficient) was selected to identify correlations between children's opinions (see section 6.2.4).

Qualitative data from the post card questionnaire (likes, dislikes, etc. relating to the workbooks) were thematically analysed following a two-step process. For example, as shown in fig 3.2, in response to the question "Why did you like it?" (referring to best activity) responses referring to The Trip activity included replies such as "You get to make it up with imagination", "We get to draw" and "Because you can make it as creative as you want" were given initial key words themes of 'drawing', 'imagination' and 'creativity'. These were then reduced to the response theme of Creativity; the children enjoyed that activity because it allowed them to be creative. Frequency tables were constructed based on the analysis highlighting preferred evaluation activities.

Best Activity	Response	Key Word	Themes
Trip	It was a bit hard and made me think	hard / think	challenge
wordsearch	because it was fun	fun	enjoyable
wordsearch	because it was a new challenge to achieve	challenge	challenge
trip	you get to make it up with imagination	imagination	creative
wordsearch	word search was just good	good	enjoyable
wordsearch	because it was hard	hard	challenge
wordsearch	because its very good	good	enjoyable
wordsearch	finding the words was good	good	enjoyable
trip	because the thing I did was funny	funny	creative
trip	we're allowed to be honest	honest	being honest / other
trip	mine shows you can say sorry and make friends	sorry/friends	making friends / other
trip	we get to draw	drawing	creative
wordsearch	it was very fun and has lots of words	fun	enjoyable
maze	liked everything	liked	enjoyable
maze	because its fun	fun	enjoyable
wordsearch	I liked finding words	liked	enjoyable

Figure 3.5: Sample of thematic analysis data

3.5 Ethics, Recruitment and Consent

All studies described in this thesis were conducted in conjunction with the eCute project and ethics approval was granted from the University of Sunderland's Ethics Board in advance of recruiting in schools.

All participants who took part in the research in this thesis were children aged 9 to 11 years old and were recruited from schools in the North East of England. Consent was gained from the parents / guardians using a consent form that was handed out by class teachers to be taken home, signed by parent / guardian and returned before children took part in any study. Both parents and children were provided with full details of each study in parent and child information sheets. Consent forms are provided in Appendix B. Parent and children information sheets are provided in Appendix C.

3.6 Summary

This chapter has outlined the research methodology to be used to explore the research question "*Are the outputs of evaluation with questionnaires improved when participant engagement informs questionnaire design?*" The research position and approach was detailed, highlighting that the research is positivist, empirical, transdisciplinary and user-centred, using Participatory Design and applying mixed methods, as detailed in the four-phase research design to inform, design, implement and evaluate the evaluation. The following chapters present the preliminary studies (chapter 4), questionnaire implementation (chapter 5) and the meta-evaluation (chapter 6).

4 PARTICIPATORY DESIGN STUDIES

With the aim of this research to increase engagement in evaluation, in this chapter, Participatory Design and evaluation studies investigating participant engagement and the reduction of response biases whilst generating quality data are explored:

4.1 Basis and Scope of Studies: Introduces the basis and scope of the preliminary studies that form this chapter. The main themes identified from the literature review are revisited and the focus of the preliminary studies is provided.

4.2 Questionnaire Workshop: Presents the participatory design approach of a questionnaire design discussion forum and workshop. In the discussion forum children were given a set of standard questionnaires. The children discussed what they liked and did not like about the questionnaires. Children then took part in a participatory design workshop in which they designed their own questionnaires on a subject of their choice.

4.3 Five Degrees: Visual Likert Scale Development: Discusses the Five Degrees Study. This section describes a series of studies in which a five-point smiley face Likert scale was developed. The aim of the study was to attempt to improve the variation within response choices provided.

4.4 Language Study: Describes a language study in which the 29 questions that form the pre- and post-test evaluation battery were tested for familiarity and understanding and then improved where necessary by children. Finally, all questions were tested for (age appropriate) readability scores using an online Flesch–Kincaid Readability Test.

4.5 Stickers as a response method: Discusses the use of stickers as a response format in order to reduce response bias, produce optimal responses and increase engagement by providing a novel experience.

4.6 Nine Square: Describes a study called Nine Square, a participatory design study that used a comic strip format as a method of increasing engagement in the provision of qualitative data by participants.

4.7 Summary of Findings: Concludes the chapter by providing a summary of findings and a discussion of how each of these findings contributed to the final evaluation study detailed in chapter 6.

4.1 Basis and Scope of Studies

The literature review identified two key themes relevant to the aims of this research:

- 1) Constructs and measures of engagement, as both relevant and feasible in terms of implementation within the scope of this research.

2) A set of response biases commonly encountered during evaluation with children. The response biases were examined in terms of a) potential causes of response biases, b) their manifestation during evaluation and c) probable harm to evaluation outputs and results.

These two themes provided the focus for a set of preliminary pilot studies investigating potential methods for reducing response bias and increasing participant engagement. The purpose of these studies was to investigate the following issues:

- Which aspects of evaluation are engaging and which are not engaging to children?
- How can high quality data be collected in a way that is engaging for children?

A range of studies were undertaken including full day workshops with a range of sessions and activities; interactive design and/or pilot evaluations of the prototype; series of interlinked one-off experiences of evaluation materials; and discussions of evaluation and materials. Table 4.1 summarises the various studies and their foci.

Study	Focus
Questionnaire Design Workshop	To understand what is or is not engaging about standard questionnaires from the perspective of a 9 to 11 year old child. A Classroom Discussion Forum and a user centred design study of questionnaires.
Five Degrees- Visual Likert Scale Development	A series of studies to develop a visual Likert scale
Language Study	A study/exercise to test and improve the language used in the 3 questionnaires that are used in the workbooks.
Engaging with questionnaires: Stickers as a response method	Investigating the use of stickers as an alternative to pen/pencil as a method of answering questions
Visual questions & answers: Nine Square	An investigation of an alternative method of collecting qualitative data.

Table 4.1: Summary of studies presented in this chapter

4.2 Questionnaire Design Workshop

The R&D requirements of the MIXER summative evaluation (see appendix A) included the use of pre- and post- test questionnaires. The questionnaires were Bryant's Empathy Scale (Bryant 1982a), the MESSY Scale (Matson et al. 1983) and the Cultural Intelligence Scale (CQS) (Ang, et al. 2007). With questionnaires being a critical element of the summative evaluation, a participatory design workshop was held with children aiming to gain a user-centred perspective of questionnaire design. The aims were:

1. To understand what is or is not engaging about standard questionnaires (e.g. a black and white, printed document as shown in the example below, figure 4.1) from the perspective of a 9 to 11 year old child. This study investigated the research question: *What do children like and dislike about questionnaires in a standard format?*
2. To gain an understanding of how a child would design a questionnaire.

The hypothesis supporting this study is therefore that questionnaire design for children will be improved if informed by the design recommendations of children themselves. The specific research questions to be answered from this study were:

- a. **What design elements would the children include in the questionnaire?** I.e. would there be a frequent use of colour? Would pictures or other visual design elements be added etc.
- b. **What would be the most frequently used response item?** Which response item i.e. multiple choice, Likert scale etc. would the children show a preference for?
- c. **What will be the narrative approach of the questionnaires?** Will the narrative be serious or humorous? Will the theme of the questionnaire be consistent or contain a variety of topics?

Bryant's Empathy Index for Children

Tick *Yes* or *No* if you agree with the sentence.

UNDERSTANDING FEELINGS (F1)

9 Girls who cry because they are happy are silly.	Yes <input type="radio"/> No <input type="radio"/>
3 Boys who cry because they are happy are silly.	Yes <input type="radio"/> No <input type="radio"/>
20 I think it is funny that some people cry during a sad movie or while reading a sad book .	Yes <input type="radio"/> No <input type="radio"/>
2 People who kiss and hug in public are silly.	Yes <input type="radio"/> No <input type="radio"/>
21 I am able to eat all my cookies even when I see someone looking at me wanting one.	Yes <input type="radio"/> No <input type="radio"/>
16 It's silly to treat dogs and cats as though they have feelings like people.	

Figure 4.1: An example section of a standard questionnaire document

4.2.1 Procedure

The aim of the workshop was to inform the pre- and post- test questionnaire design from the user perspective, with 68 children engaging in the workshop. The procedure is outlined in (see figure 4.2) and further detailed below.

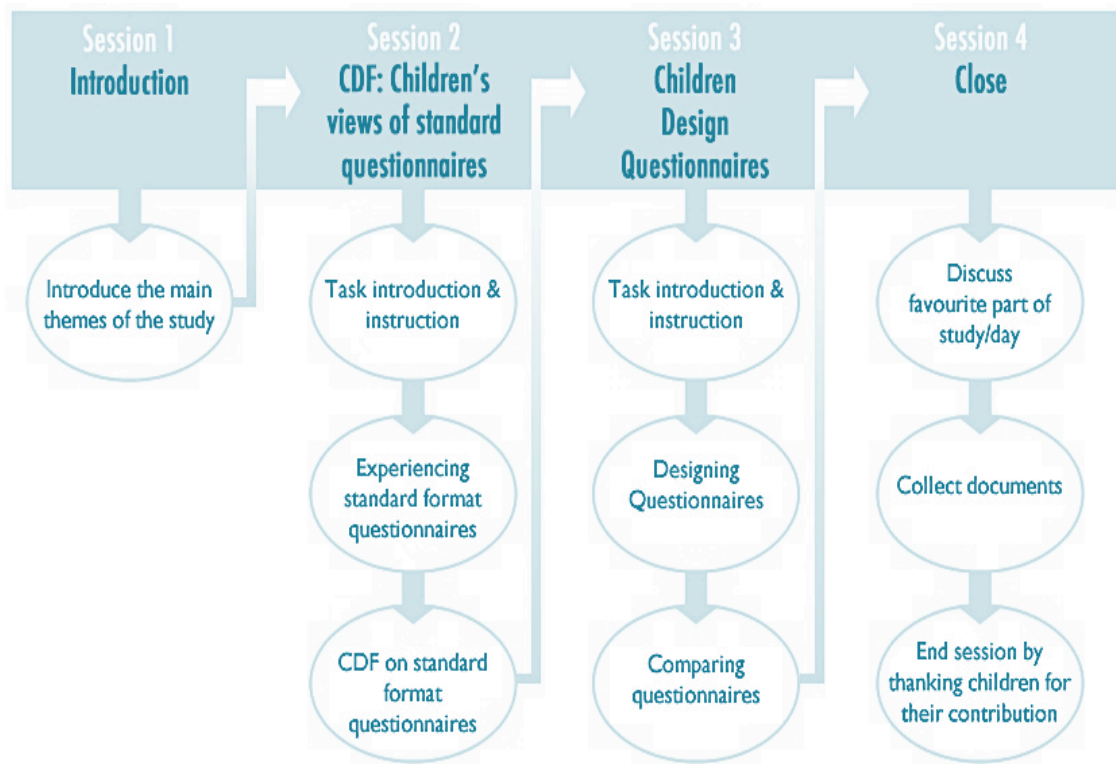


Figure 4.2: Questionnaire Design Workshop Sessions and Tasks

Session 1: Introduction

The session began by introducing a range of issues, including what researchers do, how researchers find things out and why research is important. This was followed by a short overview, explaining the activities that the children would be participating in during the workshop.

Session 2: CDF: Children's views of Standard Questionnaires

Task 1: Task Introduction & Instruction: A short overview introduced and explained the activities that the children would be participating in during the session. Children were seated at two tables in groups of eleven (three groups of around 22/23 children participated in three sessions). A teaching assistant from the school and the evaluation facilitator were also present.

Task 2: Experiencing questionnaires in standard format: Each child was provided with a set of 3 questionnaires Bryant's Empathy Index (Bryant 1982), the MESSY Scale (Matson et al. 1983) and the Cultural Intelligence Scale (CQS) (Ang, et al. 2007)). The questionnaires were presented in their standard black and white format (see figure 5.1 for an example of Bryant's Empathy Scale in standard format).

The children were given five minutes to look at, read, and answer a few questions from the questionnaires. Instructions were as follows:

"You all have the same questionnaires. Have a look at them, and read some of the questions. You don't have to answer all of the questions but try answering a few from each questionnaire. You can talk to the people in your group about what you think of the questionnaires. In about 5 minutes I'll ask some questions about what you thought of them."

There was no communication between the children and the evaluation facilitator during the first part of the task. Occasionally the teaching assistant had to intervene when children were being overly loud etc. After five minutes the classroom discussion forum commenced.

Task 3: Classroom Discussion Forum on Questionnaires: Children were asked to stop talking and to listen the facilitator while the format for the CDF was explained. In each session the CDF was explained in the same way, this was as follows:

“Now that you’ve all looked at the questionnaires I’d like to know what you all think of them. I’ll ask some questions and if you want to tell me what you think then put up your hand.”

Children kept their copies of the questionnaires to help aid any conversations that took place. With the focus of the CDF on children’s views of standard questionnaires, the discussion was prompted by asking the following questions:

1. What did the children think of the questionnaires generally?
2. Did the children like the appearance of the questionnaires?
3. Did the children understand the questions?
4. Did children understand Likert scales as response elements?
5. Which of the response elements did the children prefer?
6. What would make the questionnaires better?
7. Does anyone else have anything else to add?

The CDF was recorded using an audio recording device. Once all the questions had been discussed, the facilitator thanked the children for

their contributions concluding the session. Children were then given a short break to go outside and have a snack and a drink etc.

Session 3: Children's design of questionnaires

Task 1: Introduction: The study began by giving the children a short briefing on standard elements frequently found in questionnaires, i.e. a title, instructions, questions, response elements such as smiley Likert scales, yes or no response options, lines for free text response and boxes for drawing pictures. These were shown on a PowerPoint slide on a large screen, see figure 4.3. The examples were left on the screen for the children to refer to while they designed their questionnaires. The examples given to the children were fairly sparse, with limited colour and graphics, with the aim not to influence the children's designs.

Multiple choice

What colours do you like? Tick the ones you like:

Red

Green

Blue

People chose one or more answers

Writing

Tell me about something: _____

People write an answer or story

Drawing

Draw a picture as an answer:

People can draw a picture instead of writing

Figure 4.3: Examples of questionnaire elements for children

Task 2: Designing Questionnaires: The children were seated in two groups of ten/eleven and given colouring pens and sheets of paper. Children were instructed to design a questionnaire, on a subject of their own choice, using the elements given as examples. Children were also encouraged to be creative and invent their own questionnaire response elements if they could. Children were given 20 minutes to complete their questionnaire.

Task 3: Comparing Questionnaires: When the children finished designing, the groups switched questionnaires, so that each child completed a questionnaire from a child in the opposite group. Ten minutes was allocated to questionnaire completion. Questionnaires were then returned to the designer so they could see the answers given.

Session 4: Close

Task 1: Discussing favourite part of the day: A short discussion was then held in which children were encouraged to speak about their favourite part of the activity.

Task 2 & 3: Session end: The questionnaires were collected for analysis and the children were thanked for their contribution.

4.2.2 Questionnaire CDF - Results & Interpretation

It was clear from the beginning of the session, by observing the children, that they were not engaged by the questionnaires. The children's physical expressions displayed that they were displeased and disengaged. The children were observed to be frowning, sighing and yawning etc. As described by Hanna & Risdén, (1997), these are physiological signs of disengagement. Hanna & Risdén also prescribe that children in this age group should be capable of focusing on an activity from 30 minutes to an hour. The 5 minutes allocated to the task should not have been so taxing that the children were exhausted, however, the children were showing very obvious displays of disengagement with the materials they were given.

The CDF recordings were reviewed, a summary of responses is provided in table 4.2. The findings from this study, similar to the physical responses, are that children were disengaged and disliked the questionnaires. These results were expected and corroborate the findings of (Horton, Read, & Sim, 2011; Jensen & Skov, 2005 and Markopoulos, Read, MacFarlane, & Höysniemi, 2008). As demonstrated in table 4.2, the children responded negatively to the questionnaires. The children commented that the questions were "too hard", even though two of the questionnaires used (Bryant's Empathy Index and MESSY) were designed for use with children. Children also commented that the questionnaires looked like a SAT test. The children noted and disliked the lack of graphical media other than text. Most media designed for children contains context related imagery, stimulating colours, pictures and a variety of

fonts (see appendix H). During the CDF the children asked if they could use different coloured pens when answering the questions – demonstrating a strong desire on the part of the children to make the questionnaires more colourful and interesting. During the CDF the only positive feedback that the children provided was about the Likert scale method of answering the questions. All children agreed that they would prefer the questionnaires to be more graphical and colourful.

1 What do you think of the questionnaires?	
Rationale	Response themes
Opening with a general question gives the opportunity to give initial responses without being prompted.	"Boring", "Nothing", "Stupid", "They ask daft questions", "They were ok"
	Summary of Responses
	The majority of children's responses to the questionnaires were negative.
2 Do you like how the questionnaires look?	
Rationale	Response themes
From the review of the literature, aesthetics was identified as an important element in engagement.	"They look Boring", "Look dull", "Would be better with pictures", "They need to be more colourful", "They don't look fun", "Why don't they have pictures and colour?" "Too grown up", "It just all words", "They look a bit like the SATs"
	Summary of Responses
	The children were very vocal on the subject of the appearance of the questionnaires. With a lot of agreement from all children that the questionnaires were not visually appealing or appropriate.
3 Did you understand the questions	
Rationale	Response themes
If the children did not understand the content of the questionnaire this could cause a negative reaction and low levels of engagement.	"No", "It looks too hard", "I understood some bits."
	Summary of Responses
	The group was split between those who said they did understand the questions and those who didn't. This is reflective of reading/comprehension abilities in most classes of the age group.
4 Did you understand how to answer the questions?	
Rationale	Response themes
Did the children understand the response items that were used in the questionnaires?	"Yes"
	Summary of Responses
	The majority of children in the group understood that they were to circle a response from the Likert scale.
5 What did you think of the way you had to answer the questions?	
Rationale	Response themes
Did the children enjoy completing a scales etc.?	"Better than writing", "I ticked mine and felt like the teacher. I went tick, tick, tick...", "It was good because you could draw circles and colour them in to make it look better", "On some of the questions I couldn't decide if I was more 4 or 5"
	Summary of Responses
	The children found the Likert scale response element to be enjoyable. Again the children made reference to adding colour to the questionnaires. The children showed that they were giving considered responses in the questionnaires as they deliberated between 4 or 5 as a response.
6 What would make the questionnaires better?	
Rationale	Response themes
What changes would the children like to see, if	"Pictures", "More interesting questions", "Make it colourful"
	Summary of Responses

any?	The children again stated with a lot of agreement that the questionnaires would be better if they were improved aesthetically. Some children agreed that the questions should be more interesting.
------	--

Table 4.2: Summary of CDF responses

4.2.3 Questionnaire CDF – Impact & Recommendations

This study provided an opportunity for children to discuss the things they did or did not like about standard questionnaires. The key results and related recommendations for designing questionnaires for children were:

- Children were visibly disengaged by the questionnaires from the moment they began the task. A key recommendation is that the design of the final instruments should be dissimilar to adult correspondence or school tests. The design should be instantly recognisable as being: a) a fun activity and b) designed for children and ideally not recognisable as a questionnaire.
- As found by Mellor & Moore (2014), Haddad, King, Osmond, & Heidari, (2012) and Van Laerhoven, Van Der Zaag-Loonen, & Derkx, (2004) Likert scales were a popular response item with children and should be included in the final instrument design. One child made the comment that they completed the Likert scale by going “tick, tick, tick...” suggesting that they had not read the questions but had simply ticked the same response item for all questions, implying a level of satisficing, acquiescence and/or straight lining (Babbitt 1989) Recognising the limitations of Likert scales, but responding to positive user feedback, it

was decided to further explore how to incorporate this engaging scaling approach whilst ensuring data quality, (see section 4.3).

Children expressed a strong desire to try to improve the questionnaires by adding colour and images. Thus, instruments need to be colourful and vibrant, with multiple media formats (e.g. photos, cartoons, etc.).

4.2.4 Questionnaire Design – Results & Interpretation

The aim of the analysis was to create a summary of design recommendations that would inform the design of the final evaluation materials. To do this each questionnaire, (66 in total), produced in session 3 of the Workshop, was reviewed. Design elements and response elements were considered, with ideas and recommendations for the design of the MIXER questionnaires identified for future development. The following sections evidence the main design and questionnaire elements that were used by the children.

4.2.4.1 Design elements

The children added cover pages, used multiple colours and added decorative elements such as drawings and borders to their questionnaires.

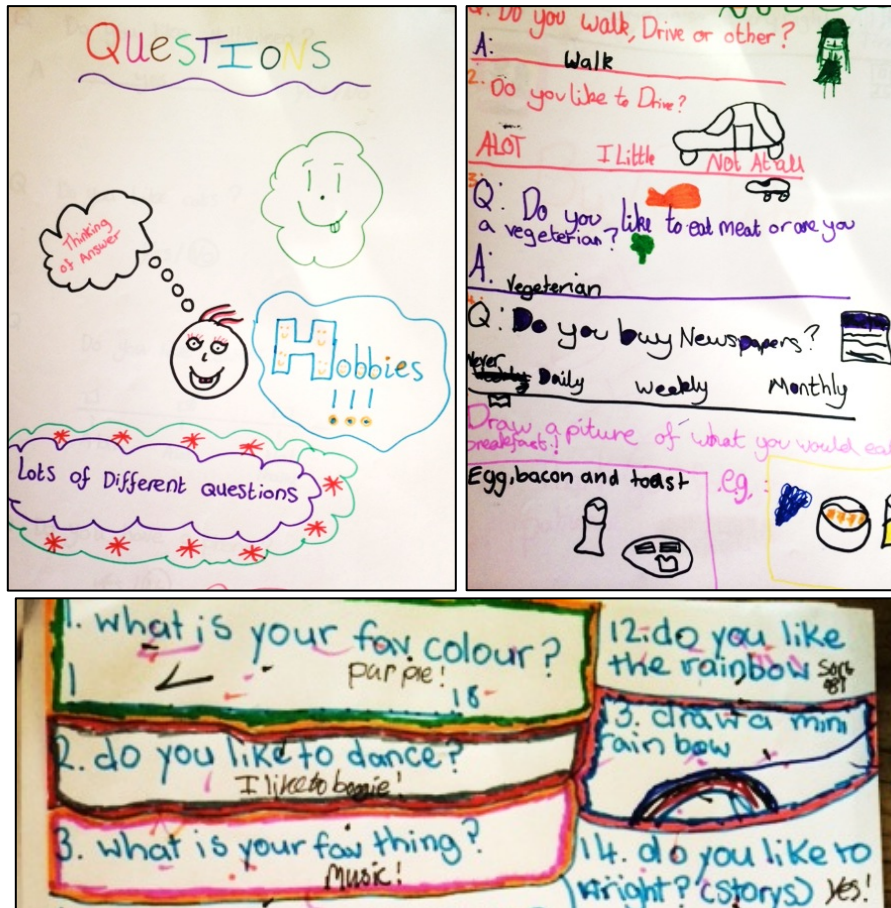


Figure 4.4: Examples of the decorative designs created by the children

It was interesting to observe that although they were not instructed to do so, the questionnaires the children produced extremely illustrative booklets, rather than single sided documents. What was particularly noticeable was the extent to which the children's designs differed to the example questionnaires they had been given in the CDF (see figure 4.1 – standard, black & white Bryant's Empathy Index).

4.2.4.2 Response Items

Some children were very creative in the design of the response items included in their questionnaires. In the example below (figure 4.5) a Likert scale is combined with a football goal in a question about football.

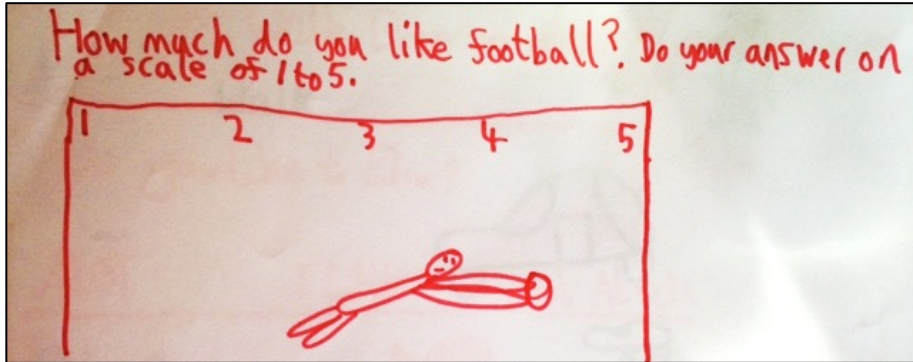


Figure 4.5: Creative use of the Likert scale format

Another child asked, "Have you ever had an event, (horrible, awesome etc.), that you will always remember" and provided a blank comic strip (figure 4.6) in which to explain the event. The child who completed the questionnaire depicted a story about their goldfish dying.

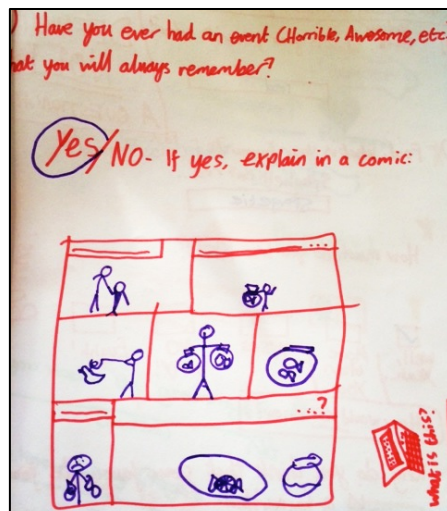


Figure 4.6: Comic strip response item

For response items children used the full range of examples given, combining Likert scales, free text and areas for drawing as a response to questions such as “Draw a picture of your favourite sport. The ‘Yes or No’ and ‘True or False’ approaches were used to achieve similar purposes in the questionnaires.

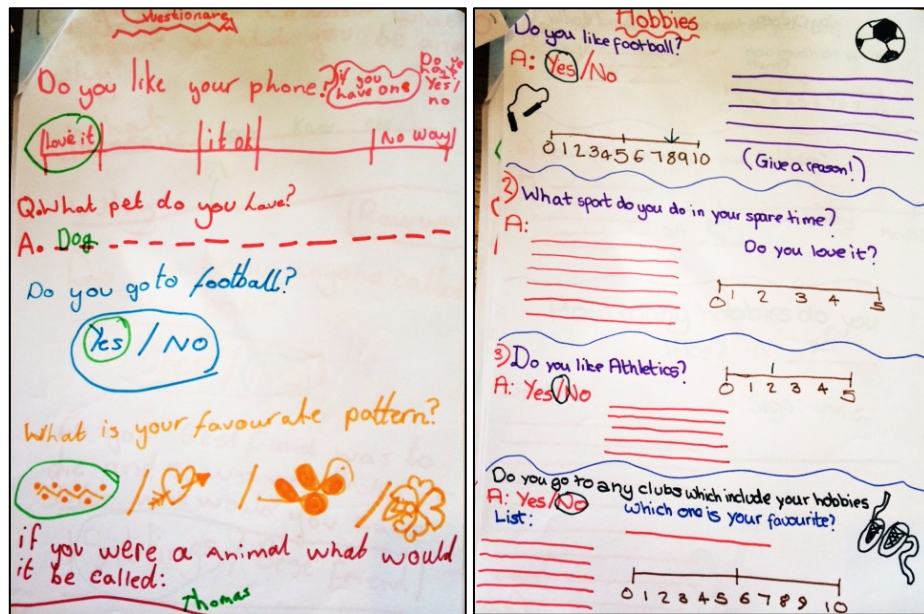


Figure 4.7: Variety in response items used in questionnaires

4.2.4.3 Narrative Approach

There was also a lot of humour in the questionnaires. During the final stage of completing the questionnaires there was a lot of laughter and talking, which contrasted with the earlier study (CDF on standard Questionnaires) when the children completed the standard questionnaires. One child added a cover to his questionnaire and the title was ‘a non-questionnaire’, (see figure 5.8). When asked why he gave it this title he explained “... *nobody would want to do a questionnaire so calling it a non-questionnaire will trick people into filling*

it out....”, further demonstrating that evaluation (in this case by questionnaire) is not something that children consider enjoyable in any sense.

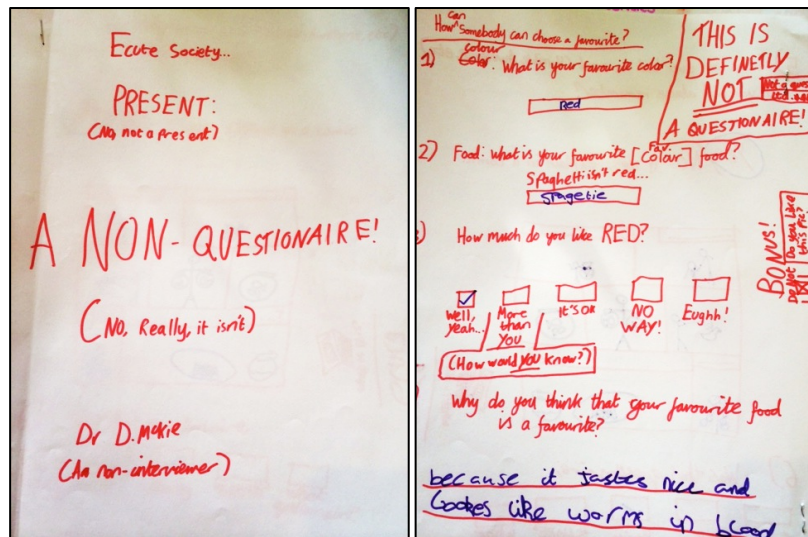


Figure 4.8: The Non-Questionnaire

At the end of the workshop the children were very eager to get their questionnaires back to see how it had been filled in and the majority of the children said that the best part of the workshop was seeing the answers the other children gave on their questionnaires.

4.2.5 Questionnaire Design - Impact & Recommendations

The children’s designs provide valuable insight into children’s expectations for an interesting, positive approach to answering questions, or quite simply, what they would like a questionnaire to look like. Through placing the child at the centre of the evaluation design as the user of that evaluation, it becomes apparent that standard questionnaires are wholly inappropriate and lacking many of the elements children expect and enjoy engaging with in hard copy

literature. The children clearly enjoyed the activities of designing and completing the questionnaires. These recommendations provided a useful reference during the design of the summative user evaluation questionnaires for the MIXER evaluation.

Recommendation	Explanation
Multiple pages	Children created 'booklets' rather than single page documents
Front covers and Title headings	Front covers had titles as seen on comics and magazines
Font variation	A variety of font presentation styles i.e size, shape, colours, patterns and decorative elements
Highly coloured	Colour combinations used throughout.
Graphical	Drawings of characters and object were frequent.
Decorative	Decorative design elements such as borders, underline, outline, dots etc.
Visual response items	Where used the Likert scales were visually creative
Combinations of response items	More than one kind of response item included.
Narrative	The narrative used in the questionnaires was informal, jovial and at times humorous.
Feedback	Children were eager to see the responses

Table 4.3: Summary of recommendations from questionnaire workshop

4.3 Five Degrees - Visual Likert Scale Development

From the workshop findings, it was evident that the children found Likert scales an engaging response item; the children enjoyed using them and included graphical Likert Scales in their own designs (see figures 4.5 and 4.7). This was a promising finding that indicated that Likert scales would engage children during the evaluation process. However, as indicated in the literature (see chapter 2) such scales can prove problematic when used with children. Children often lack the ability to optimise their response (Bell 2007) and as a result child participants in evaluation often satisfice in order to complete the task quickly (Krosnick et al. 1996), this can take the form of missing out

questions that are too taxing or ticking all of the positive responses, (Oerke & Bogner 2011).

With the findings of Mellor & Moore, (2014), Zaman, Vanden Abeele, & De Grooff, (2013) and Reynolds-Keefer, Johnson, Dickenson, & McFadden, (2009) there were concerns about how effective Likert scales actually were in gathering quality data. Thus, the aim of this series of studies was to develop a visual Likert scale for use in the final evaluation materials. The research question investigated was “What would encourage children to use the full range of available anchor points on a Likert scale to give appropriate and accurate responses?”

Four studies were undertaken, with over 140 children engaging with and assessing the MIXER interaction approach – the Pictorial Interaction Language. This was provided as an iPad application that aimed to enable communication with a character (Tom) playing an interactive game of werewolves (as discussed in appendix A). Figure 4.9 provides an early example of the questionnaire including werewolf watermark and some enhanced graphical content, e.g. colour and images etc.

Name _____
 Age _____
 Boy _____ Girl _____
 Have you used an iPad before? _____

Do you think that the Werewolves game on the iPad was:

Easy to use	☺ ☺ ☺ ☺ ☺	Not easy to use
Fun	☺ ☺ ☺ ☺ ☺	Boring
Exciting	☺ ☺ ☺ ☺ ☺	Dull
A good way to play the game	☺ ☺ ☺ ☺ ☺	A silly way to play the game
Would you have liked to play For longer	☺ ☺ ☺ ☺ ☺	For less time
Would you want to play again? Straight away	☺ ☺ ☺ ☺ ☺	Not at all
What did you think of the pictures used on the iPad? Looked great	☺ ☺ ☺ ☺ ☺	looked terrible
Easy to understand	☺ ☺ ☺ ☺ ☺	hard to understand
What did you like the most about the game?		

Figure 4.9: Early version of the PIL questionnaire

The responses of the children to the PIL are reported in Endrass, Hall, Hume, Tazzyman, & Andre, (2014) and Endrass, Hall, Hume, Tazzyman, Andre, et al., (2014). Here the focus is on the range of answers given and whether all of the five scale points are used. In addition to the study results being used to refine and improve the PIL, each study also generated results that were used as a basis to modify the five degrees of emotion shown in the Smiley Face Likerts (SFL). This multi-study procedure is further detailed below.

4.3.1 Five Degrees Procedure – Study 1: Basic SFL

60 children were asked to use two different approaches to communicate with Tom, one version was visual and icon based (see figure 4.10 – PIL) and the other was a text-based version with menus (see figure 4.11).

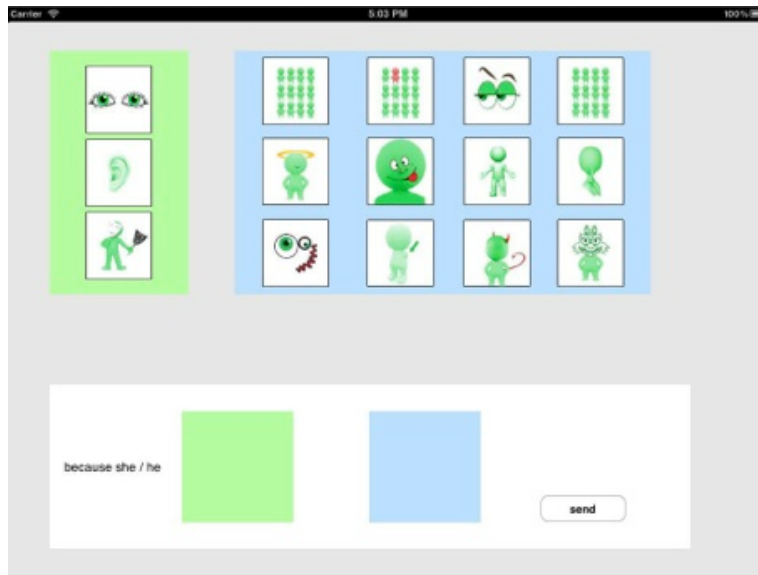


Figure 4.10: Example screen showing the Pictorial interaction language

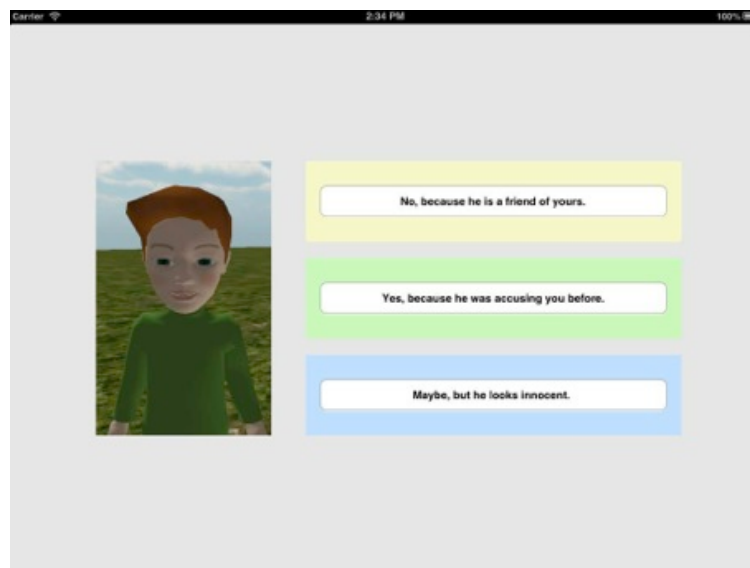


Figure 4.11: Text based version of the Pictorial Interaction Language

After using each version the children were asked to rate various aspects of the interaction on a questionnaire. Children responded by giving each aspect a rating using the scale shown in figure 4.12 below.



Figure 4.12: Initial version of the Likert Scale

In this comparative study, children were comparing systems, with one group of children using a menu/text based approach first and the other group using the PIL. Clearly, the PIL is much better and more engaging, however, for those children who used the menu/text based system first many had already selected the highest point on the questionnaire. Thus, although children preferred the PIL (this was reinforced in the discussions at the end of each session) they could only rate it the same as the previous experience.

Analysis of the children's responses identified that the children did not give a rating lower than the third/middle face. The faces shown in figure 4.13 were not used by any of the 60 children that took part in the study.



Figure 4.13: Faces unused by children to give rating responses

4.3.2 Five Degrees – Study 2: Visually appealing SFL

From the results of the first study, the initial attempt to improve responses across all the Likert scale was to improve the graphical aesthetic of the design. The scale was redesigned to make it more colourful and visual and the emotions featured were more dramatic (Reynolds-Keefer et al. 2011). Again it was found that children did not rate lower than the third face.



Figure 4.14: Likert scale used in Study 2, only the first three points were used

In addition to the final questionnaire a box was added into which the children were asked to place a sticker to indicate which version of the two they most preferred. The addition of the sticker task allowed a better understanding of each child's preference, especially for those children who rated both applications using the smelliest face, leaving no room for an improved score.

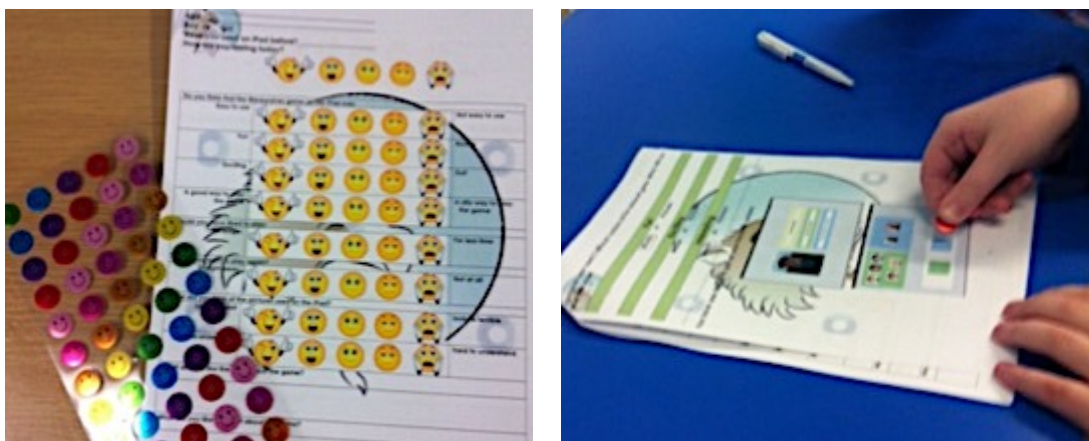


Figure 4.15: Stickers used to indicate overall preference

4.3.3 Five Degrees – Study 3: Neutral end point

In study 2, 28 children engaged with the PIL and the PIL questionnaire, with Likert scales presented as in figure 4.16. In study 3, the final anchor point was designed to show a face that was neutral rather than negative. In that interacting with the PIL is fun, the questionnaire was extended asking children to provide information about personal preferences i.e. contrasting topics such as receiving gifts and completing homework, with the aim of generating a 5. The modified version of the Likert scale as shown in figure 4.16 encouraged some of the children to rate as far as the fourth face. This was an improvement on the initial Likert scale but children still did not give any rating on any activity as a 5 (most negative face).

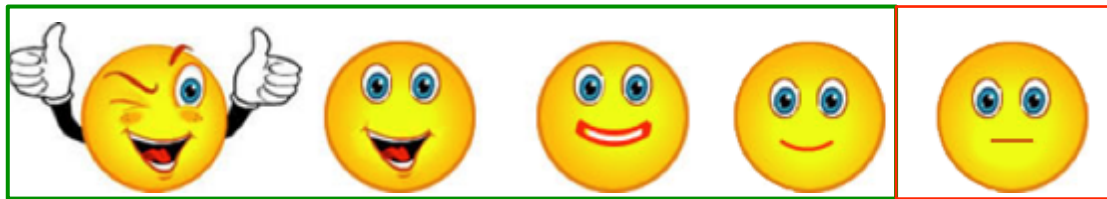


Figure 4.16: Scale with neutral end anchor

4.3.4 Five Degrees – Study 4: exploring the negative end point

The results suggested that children would not select the most negative option even if there is the attempt to inject visual humour and drama into the graphic. In study 3, the decision was taken to investigate the impact of the negative end point of the scale as a neutral face, see figure 4.17. When the negative end point was neutral, children still did not select it. In response, the end point was changed to a minimally positive face (see figure 4.17). The study was

repeated using the extended questionnaire with the modified scale. Use of this scale finally generated responses across all 5 points.

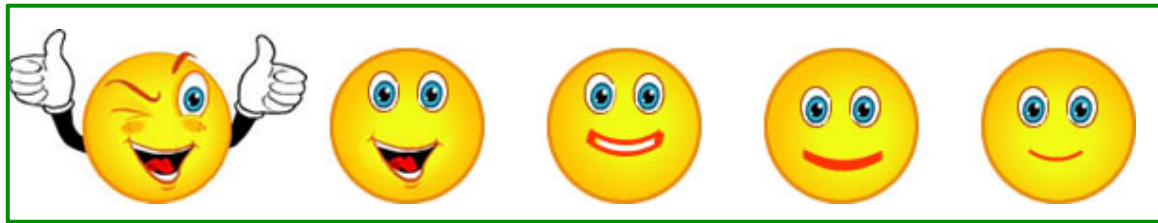


Figure 4.17: Likert scale used in Study 4

The scale in the above figure successfully encouraged more children to give more varied responses. Children responded using the entire scale, giving ratings from 1 to 5.

4.3.5 Five Degrees – Impact and Recommendations

This series of studies, further discussed in (Hall et al. 2016) identified how to improve data collection through the use of Likert scales to make it more accurate, therefore improving data quality. The main findings from the Five Degrees study that are implemented in the final study was the creation of a Likert scale that elicited a full range of responses from children, rather than extreme positive responses or responses biased by straight lining.

Although the Smiley Face Likert scale was effective, one of the findings from the review of children’s media (appendix H) was the need for variance amongst the content provided in the materials. For example, each page in a magazine designed for children used very different fonts, colours and layouts to maintain engagement by making each section novel in being different from

the previous content. This could also be seen in children's questionnaire designs.

In the evaluation context, (where the use of the scale was required repeatedly in the various questionnaires selected), there is a need for a balance of both consistency in the response approach, i.e. using the same scale, and novelty in the appearance of the content presented. In the workbooks (see chapter 6) this novelty was achieved by interspersing questionnaire pages with alternative activities that did not feature the scale.

When the scale reappeared it was used in a different layout, i.e. staggered left to right, following a curved line or placed within a maze etc. In addition to the layout of the scale the use of colour, font and the use of differing backgrounds provided a way to present the same scale in a novel and fresh way to maintain engagement with the evaluation task.

4.4 Language Study

One of the key findings from the literature review was Krosnick's (2000) work on reducing satisficing. Krosnick suggests reducing the number of questions and improving the familiarity of words used in questionnaires. The three original questionnaires selected by the R&D team (see appendix A) had a combined total of 67 questions; these were reduced down to 29 questions that were to be included in the final evaluation materials by discarding factors and questions that were not relevant to the learning outcomes of the MIXER

application. Having reduced the number of questions the second stage of Krosnick's recommendation, improving familiarity of words, was conducted in the language study described in this section.

The purpose of this study was to have children improve the language used in the questions by a) replacing any words with which they were not familiar and b) replacing any words that they felt they would not use themselves in everyday conversation/communication.

4.4.1 Procedure

66 children aged 9 to 11 took part in this three part study. For the first part of the study, the 29 questions were printed in tables on two sheets of paper. The tables had two columns, one contained the question and the other was left blank. Each child was provided with a copy of the questions and an iPad on which they could access a dictionary app and Internet. The children were asked to read the question and then mark each question as follows:

1. Underline any words that they did not know/understand
2. Circle individual words that they would not normally use and
3. Circle any questions that simply made no sense to them

Children were asked to research the unknown words using the iPad. Once they had researched the word they were asked to write a replacement for the word into the blank column. Children were asked to do the same for the words that they would not normally use. When the children had finished the questions task the second part of the study began. Part two was a group

activity in which the questions that were not understood were discussed. With the help of the evaluation facilitator, class teacher and children each of the questions was reworded on a whiteboard so that the whole group could see them and agree that they were understood by all children in the rewritten format. The third and final stage of the study was to enter each of the reworded questions into an online readability tool to check that each of the questions was of an appropriate level for 9-11 year olds; this activity was conducted separately after leaving the children.

4.4.2 Results and Interpretation

The following tables show the results of the language study:

BRYANT'S EMPATHY INDEX	
Original question	Amended question used in Workbooks
Its hard for me to see why someone else gets upset	Unchanged
People who kiss and hug in public are silly	Unchanged
Kids who have no friends probably don't want any	Unchanged
Its silly to treat cats and dogs as though they have feelings like people	Unchanged
Girls who cry because they are happy are silly & Boys who cry because they are happy are silly	Kids who cry because they are happy are silly
I think its funny that some people cry during a sad movie or while reading a sad book	Unchanged
I am able to eat all of cookies even when I see someone looking at me wanting one	I am able to eat all of sweets even when I see someone looking at me wanting one
I get mad when I see a classmate pretending to need help from the teacher all the time	Unchanged
I don't feel upset when I see a classmate being punished by a teacher for not obeying school rules. (71.8)	I feel upset if a classmate is punished for breaking the rules (81.9)

Table 4.4: Changes to Bryant's Empathy Scale following the language study

MESSY SCALE	
Original question	Amended question used in Workbooks
I call people by their names (102)	I call people by their real name (103)
I walk up to people and start a conversation	Unchanged
I show my feelings	Unchanged
I know how to make friends	Unchanged
I look at people when I talk with them (103.7)	I look at people when they are speaking (93)
I feel sorry when I hurt someone	Unchanged
I see my friends often	Unchanged
I ask questions when talking with others	Unchanged
I cheer up a friend who is hurt	Unchanged
I stick up for my friends	Unchanged
I make other people laugh	Unchanged
I share what I have with others	Unchanged
I laugh at other peoples jokes and funny stories	Unchanged
I join in games with other children	Unchanged

Table 4.5: Question changes to the MESSY Scale following the language study

CQS	
Original question	Amended question used in Workbooks
I change my verbal behavior (e.g., accent, tone) when a cross-cultural interaction requires it. (58.7)	When you meet someone new do you change the way you talk? (103)
I use pause and silence differently to suit different cross-cultural situations. (32.5)	When you meet someone new do you speak more slowly with more pauses and spaces? (95.7)
I vary the rate of my speaking when a cross-cultural situation requires it. (59.7)	When you meet someone new do you always slow down when you are speaking? (89.9)
I change my non-verbal behavior when a cross-cultural situation requires it. (44)	When you meet someone new do you change the way you move your body? (95.9)
I alter my facial expressions when a cross-cultural interaction requires it. (39.6)	When you meet someone new do you change your facial expressions? (80.3)

Table 4.6: Question changes to the CQS following the language study

The children understood the majority of words used in Bryant's Empathy Index this is not surprising as the questionnaire was designed for use with children. The only changes were to change 'girl/boys' to 'kids' and the word 'cookies' to 'sweets'. There was one question that some of the children circled; this was "I don't feel upset when I see a classmate being punished by

a teacher for not obeying school rules.” The children who circled that question explained that they found it confusing; it was therefore reworded to *“I feel upset if a classmate is punished for breaking the rules”*.

In the questions from the MESSY there were two questions that caused some confusion to the children. The first was *“I call people by their names”* the children commented, “What else would you call them?” the question was discussed and changed to *“I call people by their real name”*. The second question raised was *“I look at people when I talk with them”*, the children commented that they didn’t talk ‘with’ people, indicating that two people talking at once was not a conversation. It was suggested with agreement from the children that the question be changed to *“I look at people when they are speaking”*.

Every child circled all five of the CQS questions. This was not surprising as the CQS was the only questionnaire not developed for children and the language used is very advanced. These questions required a lot of explanation from the evaluation facilitator and the teacher about what verbal and non-verbal behaviour was in order to come up with words to replace these phrases. Similarly, the phrase cross-cultural raised a lot of issues with the children, the children had been taught in class that culture referred to different religions and/or places you may visit on holiday. While this is true, this did not reflect the view of culture(s) applied within eCute and the MIXER application. Therefore the phrase ‘cross-cultural situation’ was replaced with the phrase ‘meeting someone new’.

As a final check the reworded questions were tested for readability using an online tool (www.thewriter.com) which assesses sentence structure according to the Flesch-Kincaid readability test (Kincaid et al. 1975). The scores are included for reference in parenthesis after each of the reworded questions in tables 4.4, 4.5 and 4.6.

4.4.3 *Impact and recommendations*

The findings from this study show that when evaluating with children, it is always best (whenever possible) to use a questionnaire designed for children. This is demonstrated by the fact that none of the children understood the CQS questions.

Care should be taken when changing questions to ensure that the intention behind the question remains. There is a fine line between improving the readability and familiarity of the words in the question and writing a new question entirely.

The study tested and improved, where necessary, the familiarity of words and the readability of the 29 questions that were used for the pre- post-test evaluation battery.

4.5 Engaging Children with Questionnaires: Stickers as a response method

The main aim of this research is to increase participant engagement in evaluation; however, one of the constraints of this research was that the data collection method would have to be a paper based solution. An alternative to answering questions by pen/pencil was sought in order to a) break up the response method of activities that form the evaluation and b) as a method of slowing down those children who may succumb to acquiescence bias and proceed to “tick, tick, tick...” without taking time to provide an optimal response. From consulting age appropriate media, sticker collection books were identified as a popular pass time activity for children.

4.5.1 Procedure

The sticker method was piloted in the Five Degrees studies (see section 4.3.2) and children enjoyed this approach. After using each version of the application and answering the corresponding questionnaire the children were asked to vote for their favourite version by awarding it as the overall winner with a sticker, (see figure 4.18)

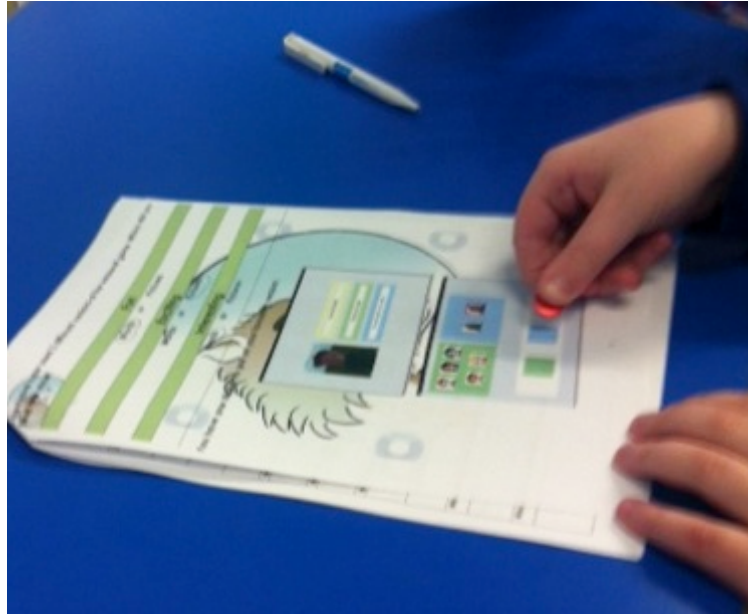


Figure 4.18: Children awarded their favourite application with a sticker

4.5.2 Result and Interpretation

A sticker as a reward system is understood by most children and using stickers to rank the best version of the two application worked very well. Novelty is one of the engagement constructs being incorporated and the inclusion of stickers provided a novel experience for the children. The children enjoyed using them and were also observed to hesitate slightly before placing the sticker on the page. This was a positive indication that stickers may provide a solution to slow down those children more inclined to rush the evaluation tasks.

4.5.3 Impact and Recommendations

Stickers were a very popular inclusion to be incorporated into the final evaluation study. The observation of the children hesitating before applying the stickers was a good indication that they could be used to slow down response reactions. This would be beneficial in questionnaire formats that offer, for example, yes or no response formats, where the likelihood of acquiescence and straight lining increase (Krosnick 2000).

4.6 Visual questions & answers: Nine Square

The Nine Square study was conceived as a method of collecting qualitative data from children. Most children prefer drawing rather than writing long passages of text. When trying to collect qualitative data this can be a problem. Having previously considered recreational media designed for children, comic/activity books were selected as an area for further investigation as a possible concept for the design of the final evaluation artefacts. Also, a comic strip was designed by one of the children in the questionnaire design workshop (see figure 4.6).

The nine squares study was conducted to test comic book layouts as a method of eliciting opinions from children on the subject of conflict resolution. The purpose of the study was to investigate if children would engage with and understand the comic strip format as a method of collecting qualitative data and whether the data collected would be useful and usable.

4.6.1 Nine Square - Procedure

36 children took part in the study, 6 groups of 6 children were each provided with the following:

- 1 x A1 sized blank comic strip
- 1 x A4 sized copy of the comic strip providing the start and end of the story, providing a very brief narrative to scaffold the story creation (Hall et al. 2012). The narrative introduced two characters, Alex and Jordan, who start out as friends, have a disagreement and are then no longer friends. The story ends with Jordan and Alex being friends again. The template containing the narrative scaffold is shown in figure 4.19.

Alex and Jordan are friends	Alex and Jordan are...	LEFT BLANK
Alex and Jordan disagree	Alex and Jordan are not friends	Something happens
LEFT BLANK	LEFT BLANK	Alex and Jordan are friends

Figure 4.19: Template provided to children

- Multiple copies of a blank cartoon styled characters to customise, cut out and stick on to the comic strip, this made the duplication of the characters throughout the comic strip easier and more fun
- Speech and thought bubbles were also provided to allow children to create a narrative

- Coloured pens, pencils, scissors and glue were also provided

4.6.2 *Nine Square – Results and Interpretation*

Each of the groups engaged with the activity and developed fully formed storylines with coherent narratives that depicted a variety of experiences on subjects ranging from sport to damaging the environment to Lady Gaga. The children were so highly engaged and enjoying the session that they complained when it was over. An example is shown in figure 4.20 below.



Figure 4.20: An example comic strip created by 'Team Pudsey'

4.6.3 Impact and Recommendation

The comic strips were viewed as highly engaging by children and they readily complete them. All groups of children produced a complete and detailed comic strip and all children collaborated with the required tasks. This study confirmed that comic book related layouts and activities are natural, familiar and enjoyable for children. The children took a long time to complete and cut out the characters, thought and speech bubbles to add to the comic strip; this delayed the completion of the task. When replicating the task for completion by an individual child, it would be advised to have children draw characters etc. onto the strip rather than cutting out and sticking the characters. This study highlights the potential of a partially completed comic strip as a data collection tool to gather qualitative data, with children clearly engaged and enjoying the format.

4.7 Summary of findings

The outcome of the studies presented in this chapter is a set of recommendations for the design of the final evaluation study. The recommendations will aim to increase engagement and reduce response bias in a large-scale evaluation study.

The questionnaire focus group identified that children were disengaged by the standard format questionnaire document and felt strongly about adding colour and decoration. The children did enjoy the Likert scale response format, but one child reported that they enjoyed going “tick, tick, tick...” when responding,

suggesting that a fully optimised response was not given and that straight lining had occurred. A solution to the straight lining problem is required. A simple solution may be to put the questions in a format that removes the linear path through the questionnaire, for example, a curved or wavy line.

The questionnaire design workshop revealed what children expect from a questionnaire in terms of visual design. For the final evaluation documents relating to this research the evaluation documents should follow the designs implemented by the children by being colourful and decorative booklets with cover pages and titles etc. In appearance the evaluation documents should look more like a comic or activity book than set of questionnaires.

The scale designed in the Five Degree study successfully elicited a full range of responses from the children. The scale will be used throughout the evaluation materials wherever a Likert scale is used in the preselected, validated questionnaires.

The stickers were very popular with the children and it appeared that the children gave greater consideration to the placement of the sticker than they would if using a pen/pencil to mark the page. The permanency of the adherence of the sticker to the page made the children pause. This was a significant indicator that stickers could be used to slow down children in the middle stage of optimising in which retrieval and formation of answers takes place. If the children can be encouraged to pause while retrieving and formulating the best answer before answering then a deeper level of

optimising has taken place, as demonstrated in the annotated version of the optimal response model shown in figure 4.21.

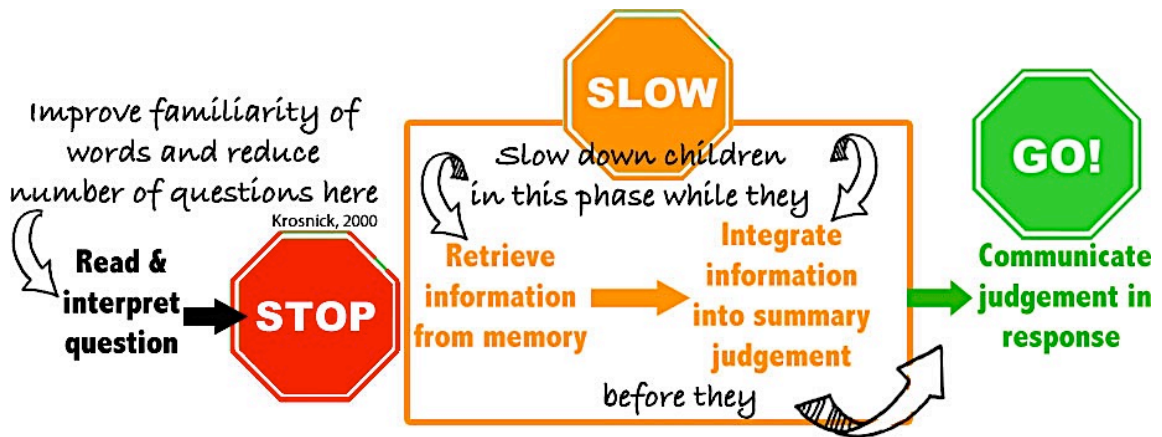


Figure 4.21: Annotated model of optimal response as applied in this research

The nine square study was designed as an alternative to the typical qualitative data collection format of a question followed by blank lines to write on as method of electing qualitative data from the children. The study was very successful and the children provided strong and clear narratives that expressed a range of views on a number of topics. The nine square/comic strip layout would work well with the comic/activity book concept and on a smaller individual scale.

The main findings from the six studies are outlined in table 4.7 below.

Study	Findings
Questionnaire focus group	<ul style="list-style-type: none"> • The Likert scales were a popular response item with children • Including colour was important to the children
Questionnaire design workshop	<ul style="list-style-type: none"> • Questionnaires should be multiple page documents with front covers and title headings – designed to be comic/activity book. • The final evaluation materials should be very visual with images and decorative elements used throughout, highly coloured. • Variety should be used in the font presentation. • Response items should be visual and a combination of response items should be used where possible. • Some sort of feedback should be included if possible.
Five Degrees	<ul style="list-style-type: none"> • This study resulted in the creation of a SFL that elicits a full range of responses that will be used in the final evaluation study
Language study	<ul style="list-style-type: none"> • The outcomes of this study were not findings as such; the study resulted in the improvement of the language used in the questionnaires. • The familiarity of the words used was improved and the CQS was reworded to an age appropriate level.
Stickers	<ul style="list-style-type: none"> • Stickers were a very popular item with the children • Further consideration was give inclusion to the study
Nine Square	<ul style="list-style-type: none"> • Comic book layouts are fun and engaging to children while providing useful qualitative data for analysis.

Table 4.7: Main findings from instrument development studies

4.8 Summary

This chapter has presented five investigative studies that explored participant engagement and response bias in user evaluation questionnaires. Through using Participatory Design and mixed methods including focus groups, user centred design, design workshops and consideration of the collection of

qualitative and quantitative data a set of recommendations was developed from the studies for the design of the user evaluation questionnaires for the MIXER Summative Evaluation. The application of the recommendations in the design of MIXER's evaluation materials is presented in the following chapter where the design of the workbooks is discussed in detail.

5 INSTRUMENT DEVELOPMENT

The previous chapter presented the preliminary studies that influenced the design of the user evaluation questionnaires. In this chapter, the approach to the design and development of the instruments as hard copy workbooks for the MIXER Summative Evaluation is provided. As detailed in this chapter the designs applied approaches to reduce biases, increase optimal answering and to increase user engagement in the evaluation process.

5.1 Instrument Development: Discusses the selection, refinement and improvement of the questions selected for the MIXER evaluation, from basic (questionnaires in their original basic form) to better (question sets reduced, language improved) to best (transformation to increase engagement and reduce response bias).

5.2 Workbook Overview: Provides an overview of the three evaluation workbooks. This includes a description of the workbook covers and user data collection. The three questionnaires as they appear (refined and transformed) as individual activities in workbooks 1 and 3. The evaluation activities for workbook 2 (EEQ) are provided along with a short description of the technical approach taken in the development of the workbooks.

5.3, 5.4 and 5.5: Describe each of the three workbooks, an overview is provided for each workbook along with a detailed page-by-page description of each activity. Questionnaire use, response bias addressed, design decisions, response methods, layouts etc. are discussed (where relevant). Images of each page/activity are also provided.

5.6 Summary: Provides a summary of the development and transformation of the instruments in the workbooks, providing a

summary of the biases and engagement constructs addressed in the design of the workbooks.

5.1 Instrument Development: Basic, Better & Best

Designing and transforming the questionnaires into workbooks was an incremental, iterative process that involved both the children (primary users of the materials) and the research team (primary users of the data). Critical R&D requirements for the MIXER evaluation (see appendix A for a more detailed description) were:

- Pre- Post- Test using 3 validated questionnaires (CQS, (Ang, et al. 2007), Bryant's Empathy Index, (Bryant 1982) and The Messy Scale (Matson et al. 1983)) to test far transfer
- In-test to assess near transfer, comprehension of the application and overall UX of the application
- Quantitative and qualitative data to be collected in a hard-copy format in the classroom situation.

Using a three-stage process to instrument development, see figure 5.1 and further detailed below, the evaluation instruments were transformed from a basic to a best evaluation experience.

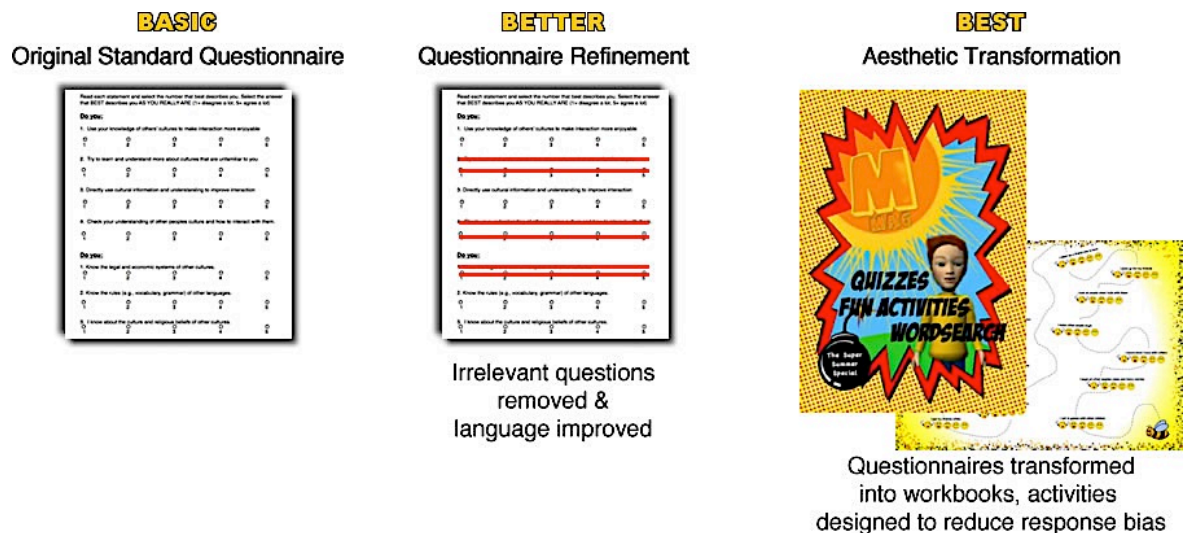


Figure 5.1: Three stage process to instrument development

5.1.1 Initial Instruments: Basic

As can be seen from figure 5.1, each instrument was initially provided in a basic format by the R&D team. For example, with questionnaires the usual approach to administering the instrument was provided, typically black and white, with numbered questions often with Likert rating scales or categories.

5.1.2 Instrument assessment and refinement: “Better”

Once the evaluation instruments have been identified, a key issue to be addressed is “how appropriate is the intended instrument and battery for the intended user group?” With many questionnaires incorporating multiple subscales or factors and possible duplication between proposed instruments, an initial evaluation of the instruments ensures only necessary data is collected.

In the initial assessment of the required far transfer instruments; a key issue was the number of questions (with 104 questions across the three

questionnaires in the pre- and post- tests), the adult focus and the repeat of the questionnaire sets.

In consultation with the R&D team, the instruments were refined, for example by only using the behavioural subscale of the CQS. With MIXER’s target age group 9-11, the comprehensibility of questions and terminology used needed consideration. Irrespective of aesthetic appeal, if questions do not make sense, seem repetitive or burdensome to answer, then user responses will be less optimal, (Larsen et al. 2008; Krosnick 2000). As only representative users can tell you if the questions are appropriate, this required piloting and several sessions with users were held to improve the language and comprehensibility of the measures (see section 4.4), with 10 questions modified across the 3 pre- post- test measures.

Instrument	No. of questions in original questionnaire	No. of questions used in MIXER Evaluation	No. of Questions improved for comprehension
CQS	20 Questions	5 Questions Behavioural sub scale	5 Questions
MESSY Scale	25 Questions	15 Questions Factor 2 Social skills /assertiveness subscale	2 Questions
Bryant’s Empathy Index	22 Questions	9 questions Factor 1 Understanding Feelings	3 Questions
EEQ	Always tailored to context so not relevant		

Table 5.1: Far Transfer - Refining the Instruments

5.1.3 Transforming the Instruments: “Best”

Inspired by child-focused hard-copy media aimed at recreational activity, such as comics, annuals and summer specials (see appendix H) and explored through Participatory Design studies (see chapter 4) the questionnaires were

designed in a comic book format. However, during piloting of the materials, children used the term workbooks. The term workbook meets children's, parent's and teacher's expectations of a classroom activity, with many schools already using workbooks in the classroom.

Three workbooks were created: for the pre-test, evaluation of immediate learning and experience of using MIXER and the post-test. As identified in tables 5.2 and 5.3, the content of Workbooks 1 and 3 was mostly predetermined by the inclusion of the CQS, Bryant's Empathy Index and the Messy Scale as part of the pre- and post- test assessment of the eCute far transfer learning goals (see appendix A), with the content of Workbook 2 provided through the EEQ (see appendix A) addressing both near transfer and user experience.

In many pre- post- test designs, identical instruments (in content and format) are used. This poses challenges in that boredom from familiarity and repetition may decrease engagement with the evaluation task. Instead, the aim was for the children to engage with the questions rather than to feel a sense of déjà vu of having done all this before in Workbook 1. Thus, Workbook 3 incorporated the same questions and instruments, but presented them within a visually very different aesthetic design, providing children with an on going, novel and engaging experience.

Care was taken to ensure that the workbooks were designed to appeal to all children. The review of children's media revealed an out dated view of boys and girls interests and design aesthetics, with pink and blue dominating the

aesthetic design of media targeted for each gender. Following the Transmedia Evaluation methodology (Hall & Hume 2011) the design of the workbooks incorporated elements that reflected the narrative of the evaluand. In this case as MIXER is set in a summer camp, the design followed an outdoors theme with grass, trees, sun, holidays and woodland animals etc. informing the design of the evaluation materials.

5.1.3.1 Lessons learnt from review of children's media

The review of children's media (see appendix H) provided useful insight into media designed for children. Clearly, the producers of children's media understand how to engage children, with valuable lessons for the designers of evaluations with children. Whilst Transmedia Evaluation (Hall et al., 2013), (see appendix A) focused purely on the obvious differences between evaluation materials and age appropriate media, with an aesthetic response to disengagement on closer inspection of children's media it is evident that designing for engagement may also provide solutions and opportunities to address the existing issues relating to response bias by improving the users engagement in evaluation. The following table provides a summary of the reviewed item, its possible purpose in reducing response bias in an evaluation context and where it was used in the design of the workbooks, along with the relevant section.

Item	Purpose	Used
Character / narrative theming throughout e.g. Doctor Who,	Introduces and reinforces character and scenario, embedding the evaluation into the narrative.	Tom is used throughout workbook 1 and 2. In workbook 1 Tom is featured on the cover page, and then in Find the camp (p. 1), Yes or No (p. 6 & 7) and on The Trip (p. 8 & 9). In workbook 2, Tom is one of the characters the children can select from the stickers for the Who Wins activity (p1), is also mentioned in the Roving Reporter (p2), True or False (p. 3 & 4), and What do you think (p. 6), Workbook-3 was testing far transfer so Tom does not appear but is named once.
Filler activities – word searches, maze etc.	The media reviewed followed up a text heavy page with either a poster, a very visual picture or an activity	The workbooks contain a variety of activities – the construct of novelty was considered and applied during the design of the workbooks.
Curved lines from McDonalds Epic Nature Spotter activity	Removing linearity from questionnaire format	Applied in all three workbooks, New Friendzzz, (workbook 1 p. 2 & 3) which woodland animal, (workbook1 p. 5) Yes or no (workbook 1 p. 6 & 7), True or False (workbook 2 p. 3 & 4)
Arrows, numbering and lines to guide readers	Used to guide from start to end of activity so that no questions are missed – ensures 100% completion with minimal interruption/assistance	New Friendzzz (workbook1 p. 2 & 3) (Workbook1) Yes or No (workbook 1 p. 6 & 7),
Comic strips	Used in media to entertain, very visual way of story telling with small amounts of text reducing cognitive effort required	The Trip (workbook 1 p. 8 & 9)
Quizzes	Engaging way of asking questions	Used in Which woodland animal are you? (workbook 1, p. 5), Epic Quiz, (workbook 3 p. 2)
Stickers	Used as an alternative to pen/pencil. May delay response and encourage optimal response	Used in workbook 2 & 3 Purposefully not used in workbook 1 to offer something new (construct of novelty) in workbook 2 Used in Who wins (workbook 2 p. 1), True or false, (workbook 2 p. 3 & 4) and in Think Fast (workbook 3 p. 4)

Table 5.2: Item of children's media, purpose in research and where used in the workbooks

5.2 Workbook Overview

All three workbooks feature a front cover with the name M-MAG (short for MIXER Magazine) and include a section for the collection of participant information, i.e. Name and age. Covers were designed to replicate the covers seen in the review of children’s media (appendix G). Each cover is shown below in figures 5.2, 5.3, and 5.4.



Figure 5.2: Workbook 1 cover and participant data page



Figure 5.3: Workbook 2 cover page

The cover of workbook 2 did not gather age or gender information as this had been collected in workbook 1. The cover of workbook 2 included a section which lead with “Today I have...” the children had just completed their first interaction with MIXER and this section gathered any immediate thoughts and feelings about the experience. The cover of workbook 2 features the words “HANDS OFF!!” the children were instructed to work alone and not to discuss or copy any answers they gave during the session. The “HANDS OFF!!” message aims to further strengthen that instruction, reinforcing the message that this is the workbook of a particular child and that nobody else should look at it as a method of ensuring that data provided is unique to each child.

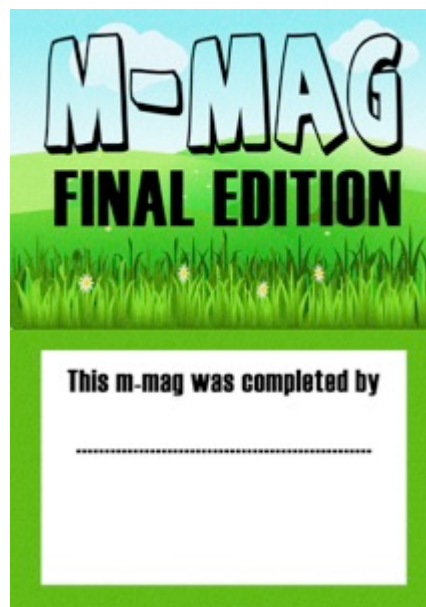


Figure 5.4: Workbook 3 cover

Tables (6.3, 6.4 and 6.5) provide a summary of the content of the workbooks, further explored in the following sections, with copies of the final versions of the workbooks provided as Appendix I, J and K.

Activity & Outline - Workbook 1 (Pre)	Measures
<p>New Friendzzz (Page 2 & 3) The 20-item MESSY is designed to look like a puzzle/maze activity, with children asked to help guide Ben to Barney. The cartoon bees are linked along a dotted line, interspersed with questions. The children move along this line 'helping' to get Ben back to Barney.</p>	MESSY
<p>Which woodland animal are you? (Page 5) Designed as a quiz, with children rating which statements are like them and which not. Children are then identified as being a Badger, Fox or Deer, where all of the possible outcomes are constructively phrased and desirable for the children.</p>	CQS
<p>YES or NO (Page 6 & 7) Presented as a comic strip, with each frame offering yes/no responses and the children following the arrow to the next box.</p>	Empathy Index

Table 5.3: Evaluation activities in Workbook 1

Activity & Outline - Workbook 3 (Post)	Measures
<p>New People, New Places (Page 1) Children are given a series of images of mobile phones and asked to text a number, 1 to 5, telling Tom what they would do when making new friends.</p>	CQS
<p>The MESSY was divided into three separate sets of questions: The Epic Quiz (Page 2) Children identify on a scale how similar/dissimilar they are to the items. Friends (Page 3) A series of questions providing learning about yourself. Maze Days (Page 5) Children make their way through a maze answering questions as they go.</p>	MESSY
<p>Think Fast (Page 4) Think fast is a sticker activity where children are provided with YES and NO stickers to use to answer the questions.</p>	Empathy Index

Table 5.4: Activities in Workbook 3

Activity & Outline	Rating Approach
<p>Who Wins? (Page 2) Having used MIXER, children should have engaged with and have a deeper relationship with Tom than any of the other characters. It was expected that the majority of children would choose to put Tom in first place. This relates to the emotional and behavioural learning objectives.</p>	Children place stickers of their 3 favourite characters onto a picture of a winner's podium.
<p>Roving Reporter (Page 3) Comprehension/opinion exercise to assess children's narrative comprehension and engagement with Tom. Higher scores for narrative show that children listened and paid attention to the story line. Positive responses equating to cognitive comprehension and deeper engagement with Tom.</p>	Varied ratings from yes/no responses, and circling correct answer
<p>True or False? (Page 4&5) Features 8 questions. 6 questions address engagement and comprehension, i.e. they have a correct true / false answer, equating to emotional, behavioural and cognitive learning. 2 questions gather children's opinions of the rule conflict reflecting the cognitive and behavioural learning outcomes.</p>	The children use 'True' or 'false' stickers to answer the questions.

<p>MIXER views (Page 5) Features questions on user experience with MIXER (e.g. appropriateness of duration, desire to use MIXER again, etc.), equating to experiential learning.</p>	<p>Children circle one of the given responses.</p>
<p>What do you think? (Page 6) Evaluates usability (e.g. voices, text, etc.) and experience (e.g. who explained the rules the best) of the MIXER application, relating to experiential learning.</p>	<p>Selections and Yes/No responses</p>
<p>iPad Page (Page 7) Provides an evaluation of the interaction approach. e.g. 'Do you think the game on the iPad was easy to use/not easy to use exciting/dull'</p>	<p>5-point Likert scale represented as faces.</p>

Table 5.5: Overview of Workbook 2 (EEQ)

5.2.1 Technical Development of the Workbooks

The pages of the workbooks were developed using adobe Fireworks (the designers preference – Photoshop or Illustrator etc. could also be used). DPI was set to 320 so that the prints would be of good quality. Each page was saved as a .jpg file. The pages were then combined into word documents (with all margins set to zero) and then saved as a PDF.

5.3 Workbook 1 – Pre-Test

The following sections outline the development of each of the activities in workbook 1 highlighting approaches taken to reduce response bias, improve optimal answering and increase user engagement in evaluation. Workbook 1 included four ‘filler activities’ these were a maze, a packing for a trip activity called ‘All packed and ready to go!’ a qualitative data collection activity called ‘The Trip’ which includes a postcard completion activity and a word search at the end of the workbook. These are classed as filler activities as they are not one of the three main measures listed above in table 6.3. These activities

were added as each session with the children lasted approximately one hour, workbook 1 needed more activities than workbook 2 or 3 as time in sessions two was taken up the MIXER interaction and session three included the classroom discussion forum.

5.3.1 Page 1 - Workbook 1 – Find the Camp



Figure 5.5: Find the camp

A maze was selected as the opening activity of workbook 1; the maze was selected as way of focusing the children and framing the workbooks as a fun activity. The maze also introduced Tom, the main character in MIXER, and shows an image of him at the start of the maze. The maze involved children helping Tom get to the summer camp (preparing the children for their interaction with Tom).

5.3.2 Page 2/3 - Workbook 1 - Messy: New Friendzzz

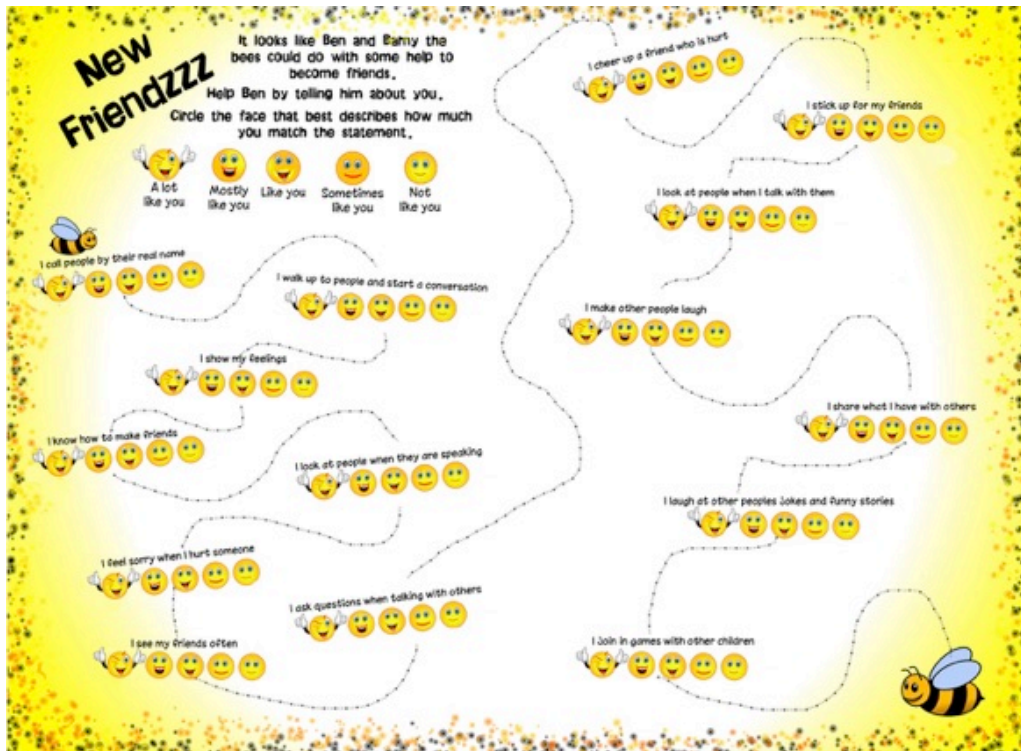


Figure 5.6: New Friendzzz

The Messy is designed to look like a maze/puzzle, with children asked to help guide Ben to Barney. The cartoon bees are linked along a dotted line, interspersed with questions. The children move along the line 'helping' to get Ben back to Barney and answering the questions as they go. The Layout of the questions, which are staggered across the page and follow a curved line, is designed to reduce straight lining. The addition of the line to follow ensures that each question is answered in turn and that no questions are missed out, this ensures full data sets and improves the quality of the data collected. The Likert scale developed in the Five Degrees study is used throughout this activity. The selection of bees as the characters in the activity sets the scene for the outdoors, summer camp narrative of the MIXER application.

5.3.3 Page 4 - Workbook 1 - All packed and ready to go!



Figure 5.7: All packed and ready to go!

'All packed and ready to go!' is short activity used to break up the sequence of data collection so that the children are not faced with page after page of questions. The activity shows items that may be taken on a trip by a child, along with an empty backpack. Children are asked to identify 5 items that they would take with them on a trip by drawing a line from the item to the backpack. This activity continues to reinforce the narrative of the MIXER experience by providing an activity that centres on preparing for a trip.

5.3.4 Page 5 - Workbook 1 – CQS – Which woodland animal?

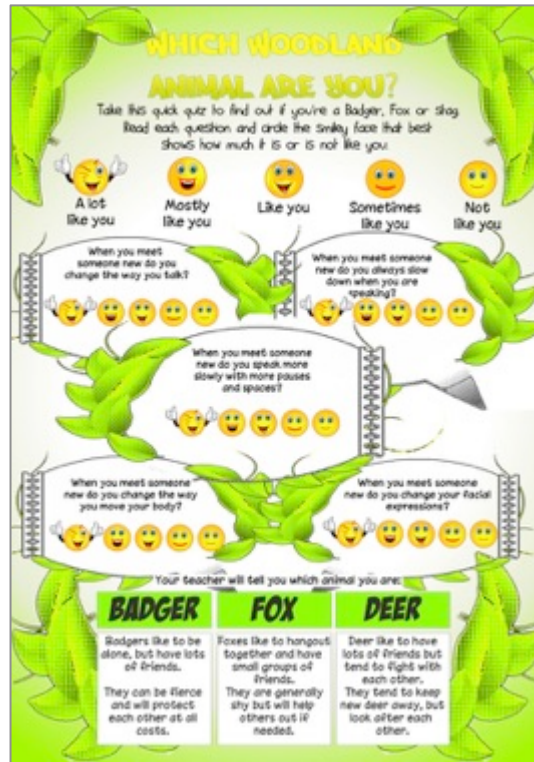


Figure 5.8: Which woodland animal are you?

The CQS was presented in a quiz format in workbook 1 as personality quizzes were identified as a popular activity in the review of children's media. Children answer the questions and are then told that their results indicate that they are a Badger, Fox or Deer. The animal is chosen at random by the evaluation facilitator. Again, the use of foliage in the design gives an outdoorsy feel to the activity, in addition to the use of woodland animals as the categories to which children are assigned. The layout of the activity attempts to reduce straight lining by placing one question in the centre of the page surrounded by the other four questions.

5.3.5 Page 6/7 - Workbook 1 - Bryant's Empathy

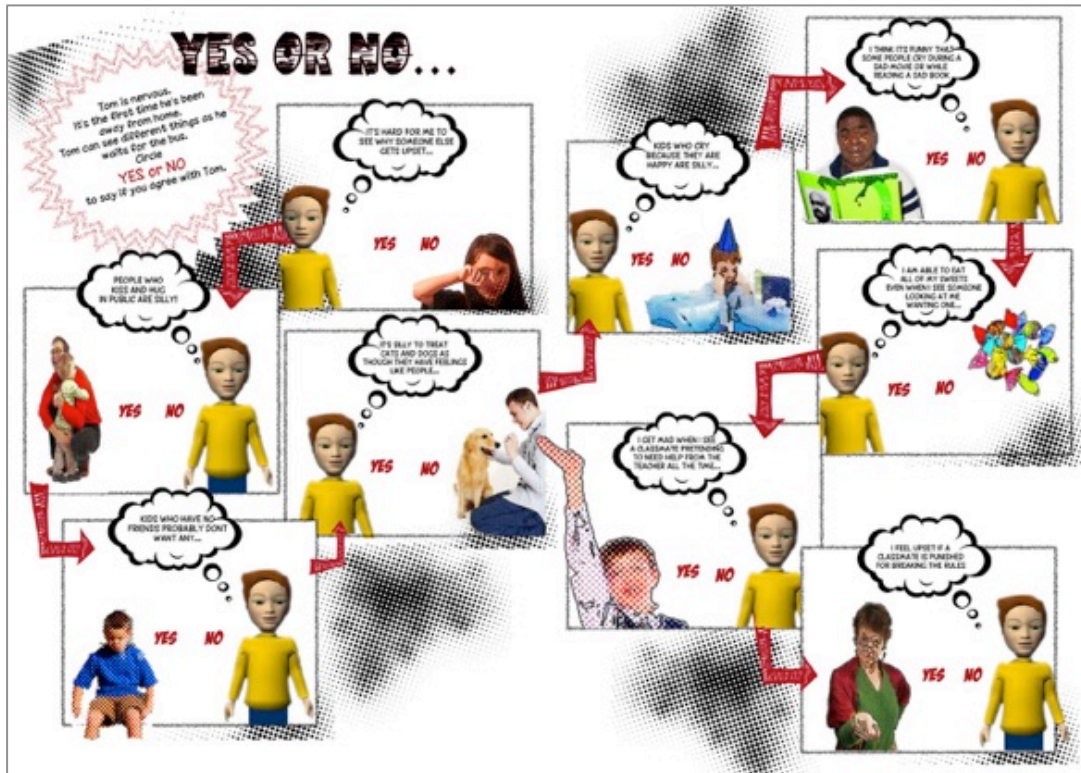


Figure 5.9: Yes or No

YES or NO is a pre test data collection activity using Bryant's Empathy Index. The activity shows each question in a box with a cartoon graphic relevant to the question, these were added to make the design more visually appealing. The questionnaire activity uses arrows to guide the child through each question, similarly to the NEW FRIENDZZZ activity this is to ensure that all questions are answered and that full data sets are collected from each child. The design was inspired by the photo story comic strips featured in many children's magazines. The comic strip shows Tom thinking, with thought bubbles that contain each of the statement from Bryant's Empathy Index, children are asked to agree or disagree with Tom by circling yes or no

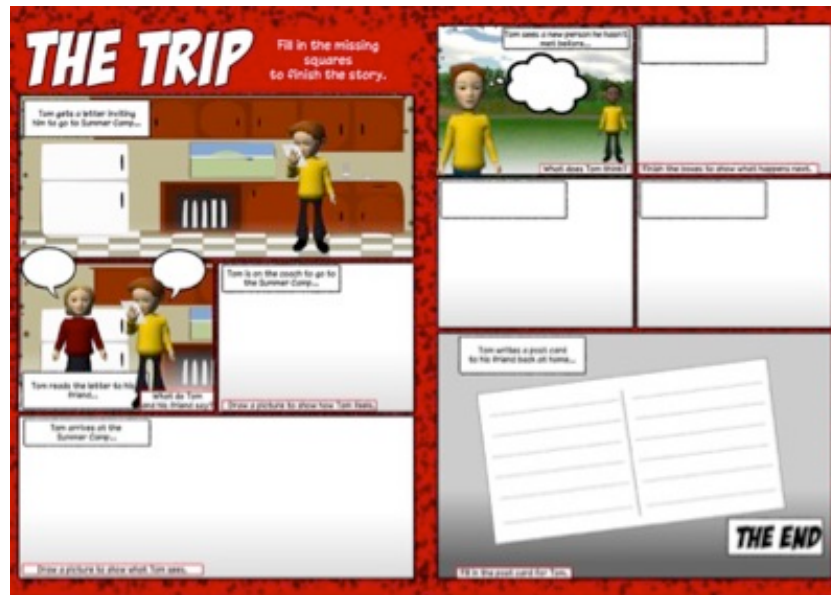


Figure 5.10: The Trip

The Trip is a comic strip activity in which the children are given half of the story of Tom being invited to go to camp, based upon the nine square study (section 4.6). In the nine square study children were given comic book elements to cut out and stick, for the sake of time in completing the workbooks children are asked to draw and write in the missing elements of the comic strip. The children are also asked write out a postcard for Tom to send home. The trip provides qualitative data on the children's perceptions of going to new places and meeting new people along with how they think another child may feel when away from home.

5.3.6 Page 10 - Workbook 1 - Summer Word Search



Figure 5.11: Summer word search

A word search was included as the final activity so that any children who finished ahead of the other children would have something to do while the rest of the class finished. The theme of the word search was summer, with flowers added as decorative elements, reflecting and reinforcing the narrative context of MIXER.

5.4 Workbook 2 - EEQ

Workbook 2, (see table 5.5), provides the EEQ – The Experience Evaluation Questionnaire. The EEQ collects data related to children’s immediate learning (near transfer); their narrative comprehension, empathic engagement; and their perspectives and views of the MIXER characters and experience. Unlike workbook 1, workbook 2 introduces the use of stickers for the children to use to answer the evaluation questions. This adds a sense of novelty to the experience by making the activities in this workbook different to the first workbook, therefore maintaining engagement throughout the evaluation tasks.

5.4.1 Page 1 – Workbook 2 - Who wins?



Figure 5.12: 'Who Wins' activity page and stickers

As workbook 2 immediately follows the interaction with MIXER it was important to link the MIXER interaction with the evaluation task as soon as possible to continue the interaction and evaluation as one experience for participants. In 'Who Wins?' Children are provided with a sheet of stickers showing every character from the MIXER application as shown in figure 6.12.

Children are asked to choose their 3 favourite characters from MIXER and place stickers onto a winners podium to show who they liked the most / first, second and third.

5.4.2 Page 2 - Workbook 2 – Roving Reporter

ROVING REPORTER

The summer camp reporter wants to write a story about Tom.
Answer the questions about Tom below.

How many games of werewolves did Tom play?
1 2 3 4 5

When you first met Tom, did he know how to play werewolves?
YES NO

Who was the best at explaining the rules? The Yellow or the Red team?
YELLOW RED

Would you want to be friends with Tom?
YES NO

Did Tom listen to what you said?
YES NO

Do you think you helped Tom?
YES NO

Was Tom... Circle one answer from each box below

Good at werewolves	OR	Poor at werewolves
Having fun	OR	Bored
Confused	OR	Knew what he was doing
Good at making new friends	OR	Poor at making new friends

Figure 5.13: Roving Reporter

The EEQ included a series of questions designed to assess the children’s comprehension of the MIXER application. In this activity questions were presented to the children in the context of a reporter who had visited the camp to write a story about Tom. The design of this activity differs to previous activities as the theme is centred on the reporter with a newspaper background and a title designed to resemble a newspaper. The reporter is shown in the centre and speech bubbles were used to show that the reporter is asking the questions about Tom. Children were provided with a multiple choice response format to answer the questions.

5.4.3 Page 3 & 4 - Workbook 2 – True or False

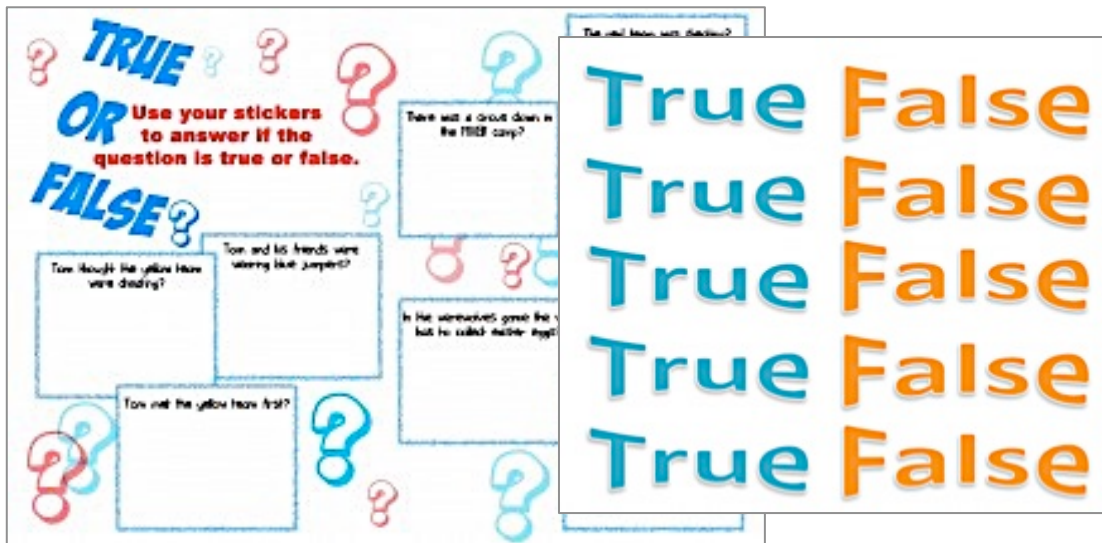


Figure 5.14: True or False activity and stickers

'True or False' is an additional comprehension study that aims to assess the child's understanding of the narrative of MIXER and their engagement with the application. The activity presents statements such as "There was a clown in the MIXER camp" that the child has to respond to by indicating if the statement is true or false. As there was not a clown in the camp, if the child answers True, then engagement and comprehension were low. Children are provided with stickers showing True or False. The alternative to using stickers would be a standard "circle yes or no" format. The aim of providing stickers was a) to provide a novel and engaging experience and b) to slow down the response reaction of the children, aiming to increase the optimal response, (Bell, 2007, Krosnick, 2000) (see figure 2.4). Children read the question, retrieve information from their memory of their interaction with MIXER, and then integrate the retrieved information into a summary judgement by choosing a sticker to add to the page to communicate their response.

5.4.4 Page 5 - Workbook 2 – MIXER



MIXER

Tell us what you thought about MIXER
by circling an answer for each question

Would you like MIXER to have lasted

Larger amount of time It was just right Shorter amount of time

Would you like to have met another group of characters in MIXER?

Yes No

How long do you think MIXER lasted?

5 minutes 10 minutes 15 minutes 30 minutes

Would you want to use MIXER again?

Yes No

Figure 5.15: MIXER activity page

MIXER' is a short activity with 4 questions about the children's general views of their interaction with the evaluand, how much they enjoyed it, would they want to use it again? Time is used as a measure of engagement in two of the questions, if the child indicates that they would have liked MIXER to last a shorter amount of time then this could indicate that time passed slowly and they were not having fun/engaged. Similarly, children are also asked to estimate how long their interaction with MIXER lasted, each interaction actually took around 15 minutes, if the child indicated 30 minutes then this may suggest low engagement, if 5 or 10 is selected then time seemed to pass more quickly indicating that the child was engrossed in the interaction.

5.4.5 Page 6 – Workbook 2 - What do you think?



Figure 5.16: What do you think?

This activity evaluates the children’s engagement with the narrative and the delivery of the narrative in MIXER. The activity asks about the rules, characters, and roles from MIXER and also asks about the voices, text and ease of understanding of the application as a whole. The design of the activity uses an eerie background which links to the night time phase of the werewolves’ game, reinforcing the link to the narrative of the evaluand. The activity combines a variety of response formats, i.e. the first two questions are answered by selecting a picture, and this is followed by a multiple-choice format where children choose one of two answers. Four Likert scale questions are also included, and finally another multiple choice of one of two answers. The various response formats are grouped to maintain some consistency, but the variety gives a comic book feel to the activity.

5.4.6 Page 7- Workbook 2 - iPad page

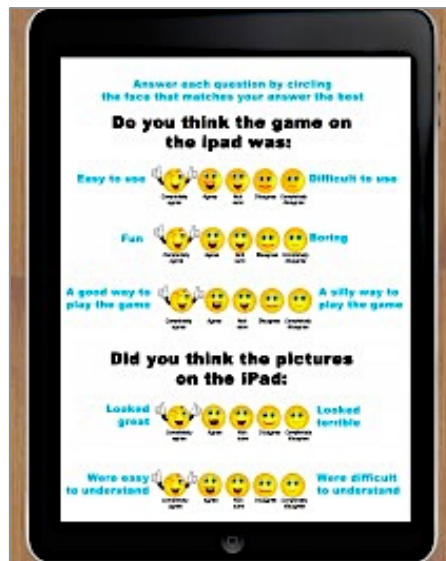


Figure 5.17: iPad page

The iPad page is designed to look like an iPad placed upon a table, this activity asks questions about the interaction modality, which was through an iPad. The visual appearance links the activity to the previous use of the iPad during the interaction with MIXER and the Pictorial Interaction Language. Again, following the transmedia evaluation methodology, this links the interaction and the evaluation into one seamless experience.

5.4.7 Page 8 - Workbook 2 - Friendship Word Search

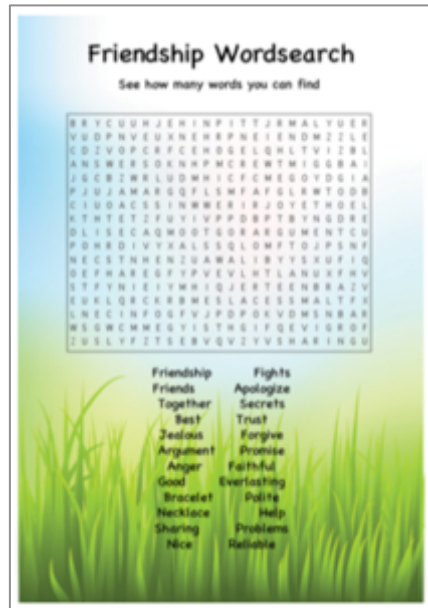


Figure 5.18: Friendship Word Search

As in workbook 1, a word search was added as the final activity so that any children who finished ahead of the other children would have something to do while the rest of the class finished. The word search used a similar outdoorsy design visual as other activities, aiming to reinforce the narrative of the MIXER application.

5.5 Workbook 3 – Post test

Similarly to workbook 1, the content of workbook 3 was predetermined by the pre/post test questionnaires. Unlike workbook 1 that prepared children for going to camp and meeting Tom or workbook 2 that used plentiful images from the evaluand, the appearance of workbook 3 was much more typical of general activity books for children. Workbook 3 did not include any of the filler type activities as seen in workbook 1, this was due to the timing schedule of the evaluation sessions. In the third session there was no interaction with MIXER, but there was a classroom discussion forum (CDF). With each of the sessions lasting for one hour, the activities in session three had to allow time to complete the workbook and the CDF. Additionally, while every effort was taken to make each of the workbooks engaging and enjoyable, by the third workbook it was anticipated that the children's enthusiasm to complete another workbook may be diminishing, therefore the workbook was kept short containing only the three pre/post questionnaires. In this workbook the MESSY Scale was divided into 3 separate activities, this further sought to reduce the similarity between the pre- and post- workbooks.

5.5.1 Page 1 – Workbook 3 - New People, New Places

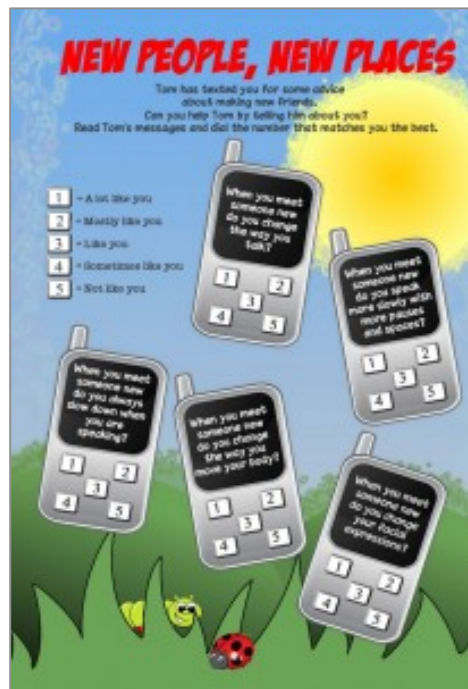


Figure 5.19: New people, new places

New People, New Places is the CQS post test data collection activity. As previously explained, workbook 3 contained fewer activities than the other two workbooks, as this was the case the use of the smiley face Likert scale was used in 3 out of 5 activities. To add variety to the activities the first page used images of mobile phones as a response item.

Children are presented with 5 images of mobile phones showing a statement on the screen, children select a number, 1 to 5, to tell Tom what they would do when making new friends. The layout of the phones and the buttons on which the children indicate their response are staggered, i.e. not linear, to reduce straight lining.

5.5.2 Page 2 – Workbook 3 - The Epic Quiz

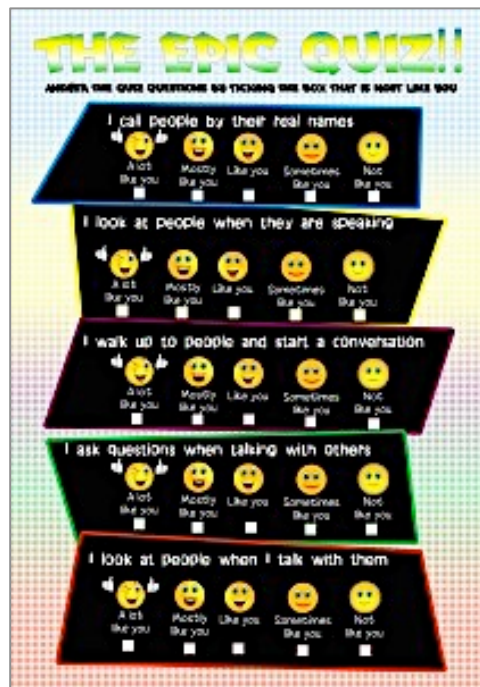


Figure 5.20: The Epic Quiz

The Epic Quiz is the post-test data collection activity, which collects part 1 of 3 of the MESSY Scale data. Designed to resemble a quiz rather than a questionnaire. A McDonalds activity sheet from the review of children’s media inspired the activity name, ‘Epic Quiz’.

5.5.3 Page 3 – Workbook 3 – Friends

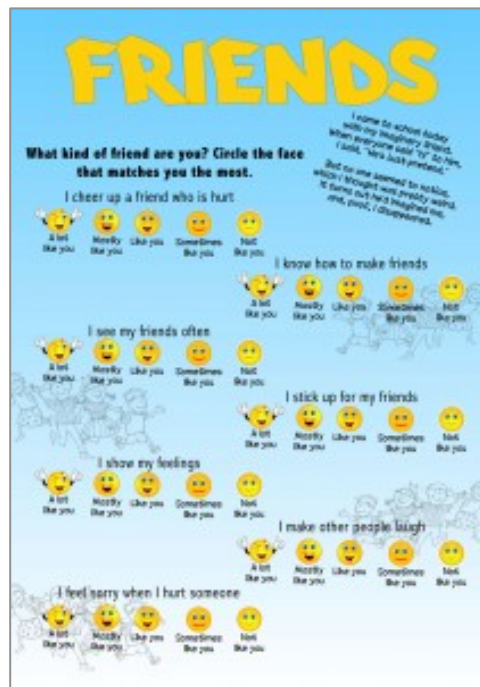


Figure 5.21: Friends

‘Friends’ is the second part of the post test data collection activity for the MESSY scale in workbook 3. The questions were positioned in a staggered layout in an attempt to reduce straight lining. A short poem about friendship was added to the page as this theming of content was seen in the review of children’s media. And it was thought that reading the poem would provide a short pause for the children before they began answering the next set of questions.

5.5.4 Page 4 – Workbook 3 - Think Fast



Figure 5.22: Think Fast

'Think Fast' is the post-test data collection of the Bryant Empathy Index. In workbook 1 Bryant was presented as the 'New Freindzzz' activity. To maintain engagement through providing a novel experience, as identified in the review of engagement in the literature review. Think Fast was designed in a very different way from its appearance and response method in workbook 1. In this version, Bryant's is presented as a sticker activity and children are provided with strips of YES and NO stickers, which are used to answer the questions. As discussed previously the use of stickers as an alternative to pen/pencils aimed to slow down the children between reading and responding to the question.

5.5.5 Page 5 – Workbook3 - Maze Days



Figure 5.23: Maze days

Maze Days combines a maze activity with the final three post test data collection questions from the Messy. The three questions are placed on the only route to lead from the start to the end of the maze. Finding their way through the maze and answering the questions as they go ensures that each of the questions is answered while providing a fun activity for the children to complete.

5.5.6 Page 6 – Workbook 3 – Spot the difference

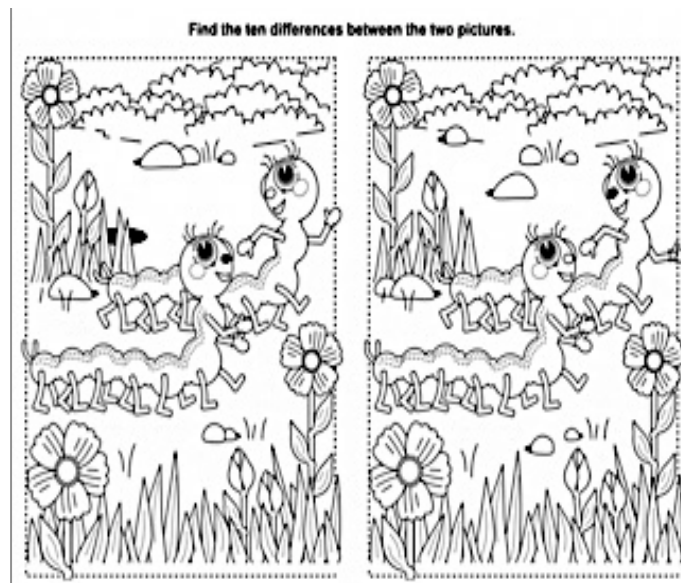


Figure 5.24: Spot the difference

A spot the difference colouring activity was added to the end of workbook 3, so that any children who finished would have something to do whilst other children in the class caught up.

5.6 Instrument Development & Transformation - Key Points

The focus of the design of the MIXER Summative Evaluation workbooks was to develop instruments that engage the users in the evaluation. Through the basic, better and best approach to instrument development, a wide range of biases and engagement techniques have been considered and designed for, as detailed in this chapter and briefly summarised here:

- **Satisficing and straight lining:** Addressed by presenting questions in

a non linear format, see Workbook 1 pages 2 & 3, 5, 6 & 7; Workbook 2 pages 3 & 4; Workbook 3 pages 1, 3, 4 and 5. The use of stickers instead of pen/pencil was also used to reduce straight lining this can be seen in workbook 2 pages 3 & 4 and Workbook 3 page 4.

- **Social desirability & Acquiescence Bias:** The redesigned Likert scale (see section 4.3) is used throughout the workbooks, see workbook 1 pages 2 & 3, 5; Workbook 2 pages 6, 7; Workbook 3 pages 2, 3, 5. Social desirability and acquiescence bias were also addressed by framing the evaluation as a fun classroom based experience with questionnaires presented as fun workbooks.
- **Novelty and fun:** Several approaches were taken to maintain the engagement construct of novelty and provide a fun experience, firstly each page applies a visually different aesthetic from the others, secondly the placement of each of the activities was such that upon completion of one evaluation task the next would provide a different activity to be completed, thirdly, the use of the stickers as a response approach also aims to provide a novel and fun experience. Finally additional activities e.g. word searches, mazes etc. were added.
- **Aesthetic appeal:** The workbooks were designed (following a review of children's media and Participatory Design) to a) appeal to the target age group b) have a gender neutral appeal c) provide a similar user experience to that of a comic book by including cover pages, combining the evaluation materials in a booklet style. This consideration of aesthetics was applied across all three workbooks.

5.7 Summary

This chapter has provided a detailed description of the design of the user evaluation questionnaires used in this research. The steps taken to improve the questionnaires from basic, to better to best were discussed. This included refinement of the questionnaires in terms of reducing question sets; the consideration and improvement of each individual question; and aesthetic considerations to increase participant engagement and reduce response biases. The following chapter provides the results of the mixed method evaluation of the MIXER user evaluation questionnaires.

6 RESULTS: Evaluating the Evaluation

This thesis addresses whether the outputs of evaluation will be improved when participant engagement informs questionnaire design. In this chapter the questionnaires developed in Chapter 5 were evaluated, providing a meta-evaluation:

6.1 Measure One: Data Quality: Data quality was assessed, through data completeness and variance in responses in individual and sample responses across all three workbooks to assess focused attention as a measure of engagement and levels response bias in the workbook data. Results and key findings are provided.

6.2 Measure Two: Postcard study: Engagement with workbook one was explored using a quantitative and qualitative postcard study, assessing fun, aesthetic appeal and returnance and what children had enjoyed most and least about the workbooks. This section describes the rationale, results, interpretation and key findings of the quantitative and qualitative data collected through a short postcard.

6.3 Measure Three: CDF: Children's responses to evaluation: Additional qualitative data was gathered in a Classroom Discussion Forum (CDF) focussing on the workbooks. This section describes the rationale for the CDF. The discussion themes, protocol, results, interpretations and key findings are presented along with observational notes taken during the CDF.

6.4 Summary: This section provides a summary of the measures and key findings presented in this chapter.

6.1 Measure One: Data Quality

The main requirement of the MIXER evaluation was to provide high quality evaluation data that enabled the eCute team to assess if MIXER had resulted in intercultural sensitivity learning (see appendix A). In this thesis the aim was to collect data in a way that engaged evaluation participants and as a result provided high quality data that was free from response bias and which provided optimal responses.

The meta-evaluation results presented in this chapter were from two perspectives: the researchers (users of the results) and the participants (users of the evaluation). In this section, the researcher's perspective is presented, with three measures were used to evaluate children's engagement with the Workbooks in relation to data quality:

Completion Rates: Focused attention, a construct of engagement, was assessed through completion rates. This assessed how complete workbook data were, i.e. how many questions the children completed. It was hypothesized that low completion rates would reveal low engagement. Low completion rates for specific questions would also highlight a lack of engagement or question comprehension.

Individual Variance: this aimed to explore how effective the reduction of response biases had been in the design of the workbooks. It was hypothesized that high variance (e.g. with children using the whole scale) would result if response biases, such as satisficing and social desirability, had been reduced. High variance would also demonstrate

high engagement, revealing that children had thought about each question and answer, thus providing a bias free and optimal response.

Sample Variance: determined if within the whole sample the entire scale had been used for each question. It was hypothesized that high variance would reveal a reduction in response biases as a result of an increase in engagement.

6.1.1 Workbook Completion and Variance: Results and Interpretation

As shown in tables 6.1 and 6.3 both workbook one and workbook three, which contained the pre- and post- test measures, had very high completion rates and variance, indicating high engagement and a reduction in response bias. Workbook two was 100% complete, as shown in table 6.2, again indicating high engagement with evaluation tasks.

Workbook Two provided the in-test measure, the Experience Evaluation Questionnaire (EEQ). This included a number of questions on comprehension where little variance was expected. Variance was expected in questions intended to measure user experience with the Smiley Face Likert scale used for two sets of questions in this workbook. Firstly, relating to the Pictorial Interaction Language, (PIL), and secondly relating to the user experience of the MIXER application. Again, with these two question sets there was considerable variance, and although most children found the PIL easy, fun

and a good way to play with MIXER, there was still considerable variance shown in the responses provided.

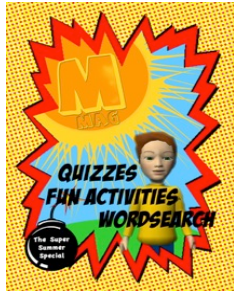
Workbook One	Questionnaire Activity	Completion	Individual Variance	Sample Variance
	MESSY	137 children (100% completion)	2 children did not use the whole range of the scale (1.46%)	98.54% showed some variability in responses (135 children)
	CQS	136 children (99.3%) 0.7% (1 child) non completion	10 children did not use the whole range of the scale (7.53%)	93% showed some variability in responses (126 children)
	Bryant	Ranged from 97.81% (134) to 100% (137) for Bryant questions	14 children did not use yes/no response scale (i.e. answered yes to all items) – 10.22%	89.78% showed some use of yes/no response format

Table 6.1: Completion rates and variance in Workbook One


Workbook Two	Questionnaire Activity	Completion	Individual Variance	Sample Variance
	User Experience of MIXER 'What do you think?' activity	132 children 100% completion	130	98.45% of children used the Likert scale range.
	Interaction/PIL Questions (iPad page)	132 children 100% completion	129	87.60% showed use of the full range of Likert scale

Table 6.2: Completion rates and variance in Workbook Two

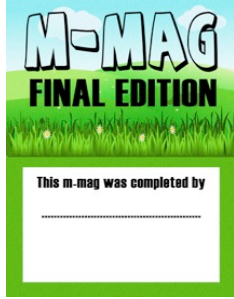
Workbook Three	Questionnaire Activity	Completion	Individual Variance	Sample Variance
	CQS	129 children at T2 (100% completion)	16 children did not use the whole range of the scale	87.60% showed use of the full range of Likert scale
	MESSY	127-129 children completed Messy at T2 (98.44% - 100% overall completion)	2 children (1.55%) did not make use of full range of the scale (answered like me a lot throughout)	98.45% of children used the Likert scale range.
	Bryant	127-129 children (98.44 – 100%) completion rate	19 children (7.75%) did not make full use of the yes/no response format and answered 'yes' to all items	92.25% of children did use the yes/no response format

Table 6.3: Completion rates and variance in Workbook three

6.1.2 Workbook Completion and Variance: Key Findings

All three workbooks had high completion rates, ranging from 97.81% to 100% complete; indicating high engagement with the workbooks and evaluation was high. Workbooks one and three also show good individual and sample variance, indicating that occurrences of response biases such as straight lining, acquiescence and satisficing were low.

The high completion rates and individual and sample variance indicate that evaluation can be designed to provide novel and enjoyable experiences that engage children. Additionally, the techniques used in this research, i.e. by increasing engagement, providing varied layouts and response formats and using the improved Likert scale etc. have shown that the impact of response

biases can be reduced. The data collected are unaffected by response biases with the majority of children using the full range of the scale provided to respond, indicating that response biases such as social desirability and satisficing had not occurred.

In addition, the data quality and use for the eCute R&D team can also be seen in a large number of publications using the MIXER evaluation data (e.g. Hall, Tazzyman, et al. 2014; Aylett et al. 2014; Lim et al. 2011; Hall, Jones, et al. 2011) and the Excellent scoring of the eCute project by the EU, where the evaluation approach and materials were highlighted as best practice and exceptionally innovative.

6.2 Measure Two: Engagement in Evaluation

The meta-evaluation considered two perspectives - that of the researchers (users of the results) as detailed in the previous section and that of the participants (users of the evaluation) detailed here. The purpose of the postcard study (see section 3.4.4) was to evaluate children's engagement with workbook one. The postcard firstly assessed general engagement using the constructs of fun, aesthetic appeal and returnance, provided as three quantitative questions (see 6.2.1). The postcard also gathered qualitative data (see 6.2.2) about the individual activities in the workbook by asking which activities were the children's most and least favourite and reasons why, and what would make the workbook better. The purpose of the qualitative

questions was to gather more detailed data about the specific activities in the workbook.

6.2.1 Quantitative data collection via postcard

Three quantitative questions aimed to assess the three areas of engagement identified from the review of engagement, these were fun, aesthetic appeal and returnance. The questions and rationale are provided in table 6.4.



Question	Construct
Was the workbook fun to do?	Fun and enjoyment , (Sharafi et al. 2006; Bartle 2004; Read et al. 2002b) were constructs of engagement identified in the literature review.
Do you think the workbook looked good?	Aesthetic appeal (Attfield, Piwowarski & Kazai 2011) was also selected as a construct to assess the children’s engagement with the workbooks. The visual approach taken in the design of the workbooks needed to be appropriate and appealing to the target age group (9-11).
Would you like another workbook to do in the future?	Returnance Read, MacFarlane, & Casey, (2002b) included the desire to repeat an experience as a construct of engagement and termed it returnance.

Table 6.4: Quantitative questions and construct assessed




The three quantitative questions, shown in figure 6.1, used the 5-point Smiley Face Likert (SFL) scale developed in the Five Degrees study (see section 5.3)

HOW DID WE DO?
 Please tell us what you think about todays session.
 Your name.....

Was the workbook fun to do?

				
Very Much	Quite a lot	It was OK	Not really	Not at all

Do you think the workbook looked good?

				
Very Much	Quite a lot	It was OK	Not really	Not at all

Would you like another workbook to do in the future?


				
Very Much	Quite a lot	It was OK	Not really	Not at all

Figure 6.1: Quantitative data collection to evaluate the workbooks

6.2.2 Qualitative data collection via postcard

The qualitative data collected via the postcard included five questions, these are shown in figure 6.2:

What was your favourite activity in the workbook?

Why did you like it the best?

What didn't you like about the workbooks?

Why didn't you like it?

What would make the workbooks better?

Figure 6.2: Qualitative data collected via the postcard

In addition to pre-test evaluation tasks, workbook one also included a variety of fun, filler activities such as a word search and a maze (see section 5.3). It was hoped that at least some children would choose an evaluation task over the more enjoyable filler activities. The qualitative questions aimed to find out what children did and did not like about the workbook and their reasons.

6.2.3 *Administering the postcard*

The postcards were handed out once all children had completed workbook one. The children kept the workbooks at hand while completing the postcard so that they could refer to the pages when answering questions. Children were asked to add their name to the postcard and to complete the postcard without discussing it with classmates. Once the postcard was completed the workbooks and postcards were collected.

6.2.4 *Postcard Results: Quantitative Results and Interpretation*

Each of the following sections will present the data and interpretation gathered from the responses to each of the three quantitative questions:

Quantitative question one: Was the workbook fun to do?

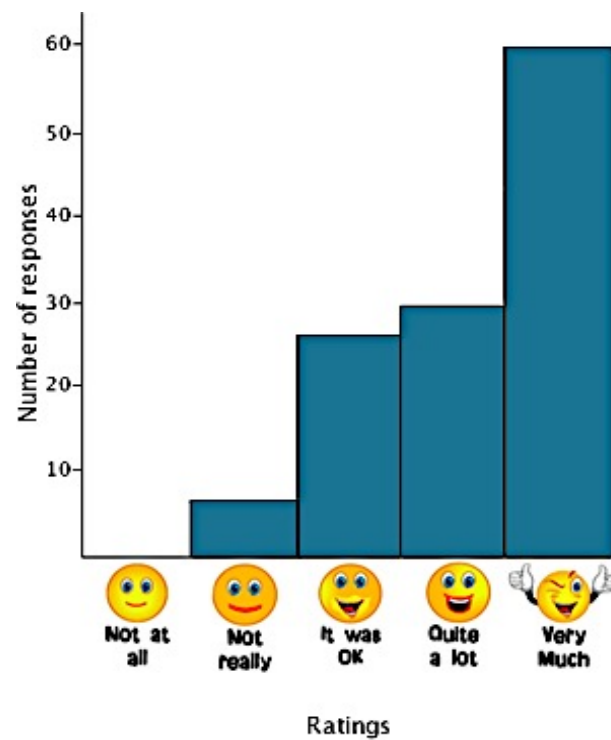


Figure 6.3: Histogram – Was the workbook fun to do (N=118)?

Figure 6.3 identifies that children had clearly enjoyed the workbook with almost 74% (n=87) of children responding positively (3 or above), rating the workbook as being better than ok. Half of the children (n=59) found the workbook “Very much” fun to do, identifying that for them it had been a fun, and thus an engaging activity. No one had disliked the evaluation experience sufficiently to rate that the workbook had not been fun at all. The mean response was 4.19 (SD= .942) indicating that children responded positively.

Quantitative question two: Did the workbook look good?

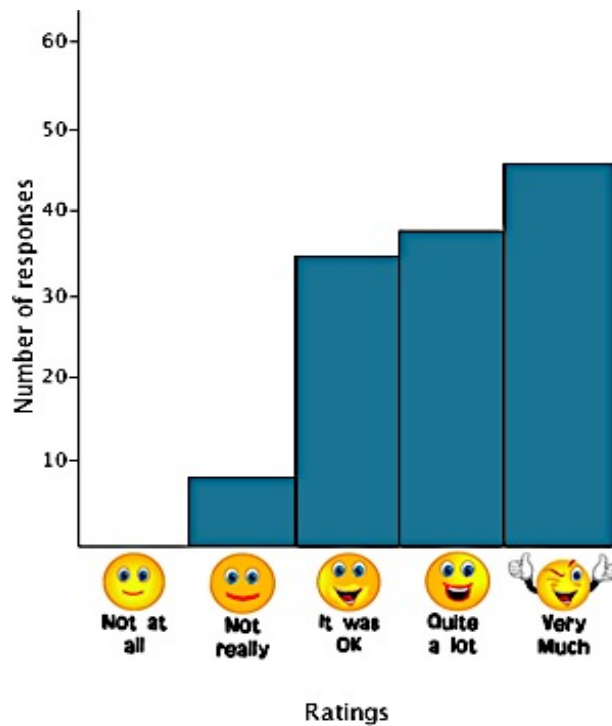


Figure 6.4: Histogram – Do you think the workbook looked good? (N=119)

Whilst half of the children had found the workbooks “Very much” fun to do, only 35.3% (n=42) agreed “Very much” that it looked good. As can be seen from figure 6.4 this distribution is less positive than that achieved for both the fun children experienced with the workbook and their desire to repeat with another workbook. However, even so, 65.6% (n=78) of children agreed that the workbook looked better than ok.

Quantitative question three: Would you like to do another workbook in the future?

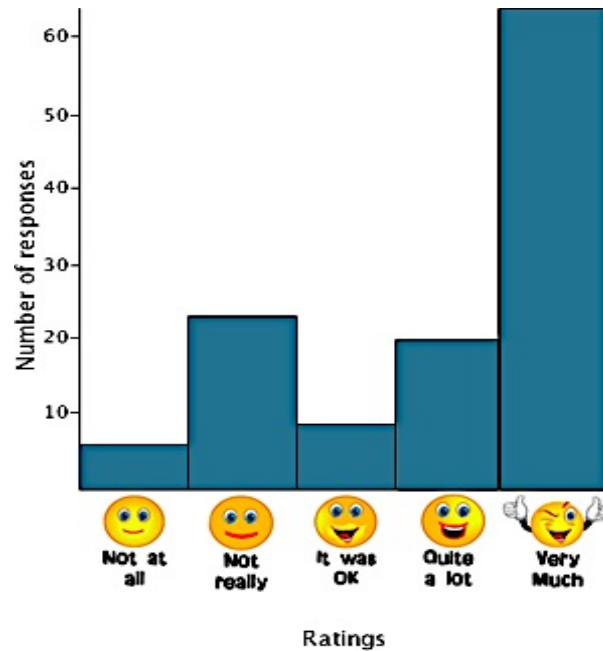


Figure 6.5: Histogram – Would you like another workbook to do in the future? (N=119)

70.6% (n=92) of children gave a positive indication that they would like to repeat their evaluation experience, with over half (53.8%) (n=64) giving the highest possible response (very much) in response. The mean response was 3.97 (SD=1.32) indicating that most children were positive about wanting to do another workbook in the future.

Due to non-normally distributed data, with most results positive, the non-parametric Spearman's rho was used to assess the whether there was an association between the three postcard questions with significant positive correlations found between the following:

- The workbook being ‘*fun to do*’ and the workbook ‘*looking good*’ ($r_s(118) = .51, p < .001$);
- The workbook being ‘*fun to do*’ and wanting to ‘*do another*’ workbook in the future ($r_s(118) = .77, p < .001$);
- The workbook ‘*looking good*’ and wanting to ‘*do another*’ workbook in the future ($r_s(119) = .52, p < .001$).

These results demonstrate that more children will want to complete another workbook (the engagement principle of returnance (der Sluis et al. 2015) if they found the workbook to be fun and if it looked good.

	Fun	Looked Good	Do another
Fun	*	.505**	.769**
Looked Good		*	.523**
Do another			*

** . Correlation is significant at the 0.01 level (2-tailed).

Table 6.4: Correlation between children’s postcard views

6.2.4.1 Postcard Variance

As detailed in table 6.5, children can be seen to be answering across the scale used on the postcard, further evidencing that the redesigned Likert scale (see section 5.3) encourages children to use the full range of the scale, with no indication of acquiescence, straight lining or extreme responses.

	Min	Max	Mean	Std. Dev.
Was the workbook fun to do?	2	5	4.19	0.94
Would you like to do another?	1	5	3.97	1.31
Did the workbook look good?	2	5	3.95	0.93

Table 6.5: Distribution of children’s responses from the postcard study questions.

6.2.5 Postcard Results: Qualitative Results and Interpretation

Workbook one contained 10 activities in total, including evaluation instruments (e.g. completing the CQS, providing qualitative data for The Trip, etc.) and filler activities that aimed to reinforce the MIXER narrative of a trip to a camp (e.g. summer themed word search, 'find the camp' maze in addition to themed styling/decorative elements throughout).

Responses to the qualitative questions were analysed using thematic analysis, this is further detailed in chapter 3, (section 3.4) with responses to each of the questions discussed below.

Table 6.6 summarises children's favourite activities, identifying that 72 (61%) children chose a non-evaluation activity as their favourite and 46 (38.6%) children chose an evaluation activity as their favourite. The word search was the children's favourite activity, which is unsurprising, as this type of puzzle is well known amongst children of this age group and it is easy and fun to do. However, the results also clearly indicate that some children preferred evaluation related activities, (e.g. a transformed quantitative questionnaire or qualitative data collection activity such as The Trip,) to non-evaluation activities.

Activity	Frequency	Valid Percent
Wordsearch	65	54.6
The Trip	25	21.0
Woodland animal	6	5.0
Maze	5	4.2
New friends	5	4.2
All Packed and ready to go!	4	3.4
Questions	4	3.4
Yes or No	2	1.7
Smiley faces	1	.8
Postcard	1	.8
Circling the faces	1	.8
Total	119	100.0

Table 6.6: Sample data from SPSS Output file for favourite activity

5.0% (n=6) of children said that they liked the ‘Which woodland animal are you?’ activity the best. Stating that, “It was the most interesting”, “You could find out what animal you are”, “It was more interesting than the rest”. These are significant results as this activity presented the CQS, the only one of three questionnaires selected by eCute that was originally for adults, and the questionnaire that children found most difficulty with in the language study (see section 4.4). During the session this activity enthused the children a lot, they were very excited to find out which animal group they were in. The addition of the interactive element of receiving immediate feedback, in assigning each child (randomly) to a group further engaged the children in the evaluation activity.

1.7% (n=2) of children stated that ‘Yes or No?’ was their favourite activity. ‘Yes or No?’ presented Bryant’s Empathy Index as a comic book styled activity in which children followed arrows to lead them to the next question.

The reasons provided were, "*It was easy*", and "*It was fun*". In this activity efforts were made to present the questionnaire in a fun and vibrant way with arrows to guide the children through each of the nine questions to ensure that none were missed.

One children said that their favourite activity was the "*smiley faces*", responses do not give a clear indication as to which quantitative data activity the children were referring to, but as the Likert scale was only used in evaluation activities this indicates that these children also found the evaluation elements of the workbook to be more fun than non evaluation activities.

Children were then asked to explain why they had chosen their favourite activity. After the word search, 21% (n=25) of children indicated that 'The Trip' was their favourite. The reasons given by the children referenced the creative elements of the activity stating they liked it because, "*We get to draw*", "*I like drawing*", "*Drawing the pictures was cool*". Other children also enjoyed the combination of activities provided in The Trip, "*It was good because you got to draw and write*". Some children provided feedback that showed a sense of pride in their work, for example, "*Mine was funny and entertaining*", "*Because the thing I did was funny*" and "*Mine was really good!*" One child provided a more sensitive response that related to the eCute theme of overcoming difference, stating, "*Mine shows you can always say sorry and make new friends*". Other reasons given were, "*It was really fun*" and "*I liked that part*", indicating a preference for the quantitative data collection activity.

As a qualitative data collection exercise, the results referring to The Trip activity were very encouraging. There is a perception that the collection of qualitative data (from both adults and children) is more difficult than the quantitative alternatives (Creswell 2012). The fact that some children a) chose a qualitative data collection activity as their favourite over the other evaluation related activities and b) expressed valid reasons for their selection, such as allowing for creativity, fun and enjoyment indicates that this evaluation activity was an engaging method of collecting qualitative data.

Interestingly, when the children were asked to indicate which activities they did not like, 31.1% (n=37) selected The Trip as their least favourite. However, reasons given for not liking this activity indicated a lack of ability, such as “*I can't draw*”, “*I'm not good at drawing*” indicating that the lack of enjoyment came from an absence of ability rather than the activity being badly designed.

One child indicated that they did not like the New Friendzzz activity, stating, “*There were too many questions to read*”.

14.4% (n=14) of children selected non-evaluation elements as their least favourite, giving reasons such as “*The characters looked scary*” and “*the maze was too easy*”.

36.9% (n=44) of children indicated that they “*Liked everything*” in response to the question “What didn't you like about the workbooks?” This could be that the children genuinely enjoyed everything or could be an occurrence of the response bias of social desirability as this was the first and only time children

were asked to give the evaluators direct feedback that inferred a negative connotation. 3.4% (n=4) of children stated that thing they didn't like about the workbooks was that they were "too short", indicating that they would have liked to have done more evaluation!

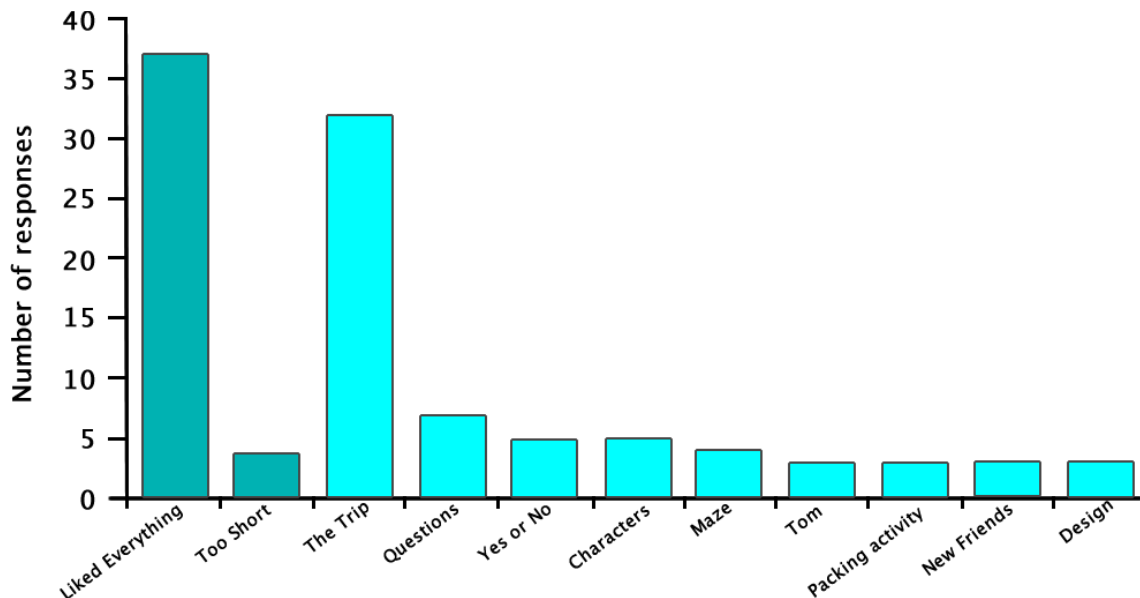


Figure 6.6: What the children didn't like about the workbooks – responses of children who didn't dislike anything are shown first in a darker shade

In response to the question "Why didn't you like it?" the most frequent response given was "I liked everything" with 37% (n=44) of children providing this response, a further two children also stated 'it was too short', indicating that they would have liked the evaluation to have lasted longer. Other frequent (more than 10%) reasons for the disliking of the activity/element named in the previous question included:

1. 19.3% (n=23) stated that they didn't like/were not good at drawing or writing
2. 13.4% (n=16) gave reasons such as "*it was boring*", "*It didn't interest me*"
3. 11.7% (n=14) children commented that they didn't like The Trip because of the characters. Reasons included, "*The kids looked freaky*", "*They looked like fakes*" and "*The characters looked spooky*". The characters as they appeared in the workbook graphics were taken from screen shots of the MIXER application. While these responses are not relevant to the children's engagement with the evaluation, it is useful feedback for the R&D team.

Other responses included "*too many questions*" 5% (6); the workbooks were hard 3.4% (4); confusing 3.4% (4) and silly 0.8% (1). These were heavily outweighed by those children who couldn't find anything they disliked, or wanted more evaluation/workbooks or dislikes due to feelings of a personal lack of ability such as lack of artistic talent. 31.1 % (n=37) of children indicated that the trip was their least favourite activity, the most frequent reasons given for disliking were 19.3% (n=23) not good at drawing or writing and 11.7% (n=14) didn't like the characters shown in The Trip (taken from the MIXER application).

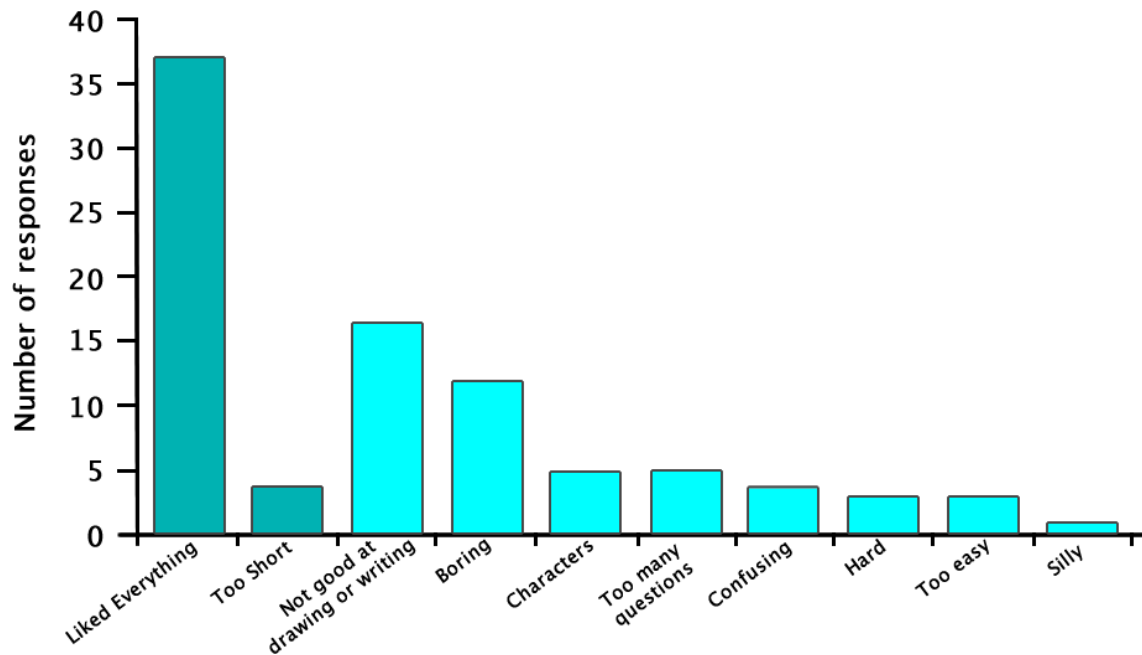


Figure 6.7: Reasons for disliking (with positive responses, i.e. there was nothing they disliked, shown in darker green).

6.2.6 Postcard Study: Key findings

The key findings from the postcard can be summarised as:

- The strongest correlation with returnance was fun, indicating that if children thought the workbook was fun they were more likely to want to do another workbook.
- Children rated the workbook lower in aesthetic appeal than fun but still indicated that they would like to do another. This was interpreted as an outcome of the design approach taken when designing the workbooks, (as informed by the review of children’s media) being correct, thus producing a workbook that was aesthetically appropriate for the age group, and therefore was not considered by the children to be anything other than normal or typical. This was corroborated in the CDF with children agreeing that the workbooks were like activity books they have at home.

- 38.6% (n=46) of children selected an evaluation activity as their favourite with The Trip activity selected the most, followed by the 'Which woodland animal are you?' activity.
- When asked what the children did not like and why, the largest response indicated that there was nothing to dislike. The second largest response (19.3%, n=23) indicated that children did not like certain activities due to lack of personal ability such as drawing.

6.3 Measure Three: CDF - Assessment of Engagement with Evaluation

The classroom discussion forum (CDF) was selected as an alternative to simply repeating the postcard data collection activity and to gather richer data with the inclusion of qualitative data. Additionally, the CDF was included for the following reasons:

- It was hoped that the discussion forum format would allow children to communicate their thoughts about the workbooks with more ease than if they had been asked to write down their opinions in a questionnaire
- The group conversation style of the forum would lead to deeper understanding, as children provide their input this may in turn prompt additional questions that may not have thought of during the development of a questionnaire or interview script.

The CDF was conducted after workbook three as the final evaluation activity to collect additional qualitative data by discussing the children's views on the workbooks. The questions asked in regard to the evaluation workbooks were as follows:

1. What did you think of the workbooks?
2. What did you like the best about the workbooks?

3. What did you not like about the workbooks?
4. Are they like comic/activity books you have at home?
5. Would you like to do another one?

6.3.1 Conducting the CDF

Children had completed workbook three and all workbooks were collected. Children remained seated at their usual desks while the evaluation facilitator explained the format and 'rules' of the CDF. The rules were that the facilitator would ask a question and the children would put up their hands to answer, the facilitator would then indicate whose turn it was to speak. This ensured that children were not all speaking at once, talking over each other etc. If the children thought of something to say while another child was speaking they were to raise their hand again. The facilitator continued to allow each child to speak until the children agreed that they had nothing more to add on that particular subject/question theme. The CDF was recorded using an audio recording app on a mobile phone. This was then transcribed for analysis.

6.3.2 CDF: Results and interpretation

Sample responses given during the CDF are provided in table 6.7.

Question/Theme	Responses
What did you think of the workbooks?	"Awesome!" – a lot of agreement.
	"They were really good."
What did you like the best about the comics/activity books?	"Making up the story in the first workbook and using the stickers in the others."
	The stickers were the best thing of all!" full class agreement
	"The word search and spot the difference activities at the end of the workbooks were good too."
	"The comics are a lot about you, it would make it a bit different if the workbooks were more about other people and how other people treat you."
	"It was good to have to think about myself and what I'm like. I never thought about myself and these things like this before."
	"I liked that I was a fox"
What didn't you like about the workbooks?	"Nothing"
	"I liked everything"
	"They were kind of the same. Some questions were in the first one and the last one." Around half (16) of the children agreed with this statement.
	"I didn't like that mine with the drawings didn't look as good as X's" With a lot of agreement that child X was the best artist in the class.
Are the comics/activity books like the type you have at home?	"Yes." Full class agreement
	"Mine at home have more different stories in them to read."
Would you like to do another one?	"Yes", full class agreement.
	"We want more stickers!"
	"More quizzes where you find out what you are"
	"Yes, with more drawing"

Table 6.7: CDF response themes

Observational data noted during the CDF highlighted that children were very expressive when discussing the workbooks. As they had been asked to only speak when putting up their hand etc., when they were not selected to speak they displayed their agreement by becoming very animated, by nodding in an exaggerated way, standing up and raising their hands as high as they could.

When the stickers were mentioned there was a lot of excitement and agreement. The children reacted similarly when discussing the 'Which woodland animal are you?' (CQS) with all children becoming excited that they were a particular animal, the same animal as friend etc. This continued during the break when the children played games as foxes and deer etc.

6.3.3 CDF: Key Findings

The CDF identified the following:

- Children enjoyed the workbooks. They also liked the stickers very much and the "Which woodland animal are you" activity.
- Children engaged with the workbooks. While considerable effort was made to create a novel experience with each workbook, some children did mention that "some" of the questions were the same. While this is a dislike from the children, it does show that those children who noticed this had engaged with the evaluation materials (by reading the questions fully) enough to recognise that the questions were repeated.
- All children were so engaged by the evaluation that they expressed a desire to repeat the experience and do another workbook
- The children stated that the workbooks were similar to the activity books they use recreationally. This may corroborate the previous interpretation that the workbooks appearance was unremarkable to the children as they are similar to everyday items used by the children.

6.4 SUMMARY

This chapter has presented the results of the three measure approach taken in the meta-evaluation that forms part of this research. The first measure assessed data quality in terms of data completion and variance in responses. The data was considered to be of high quality as the data was both largely complete and showed variance indicating that possible issues of response bias were improved by the design of the workbooks and the improved SFL scale. The assessment of children's engagement gathered in the postcard study showed that children found the workbooks to be fun and an experience they would like to repeat. The final measure reported in this chapter was the CDF session. Again children responded positively about the workbooks and were particularly positive about the stickers used in workbooks two and three and the 'Which animal are you?' activity. The next chapter will discuss and synthesise the findings presented in this chapter.

7 DISCUSSION

This research sought to answer the following research question, “*Are the outputs of evaluation with questionnaires improved when participant engagement informs questionnaire design?*”

In this chapter, the approach to answering this research question is discussed, considering the approach to investigating if evaluation can be designed to both engage participants and provide high quality data by reducing occurrences of response bias. The key findings as presented in each chapter are further considered along with a discussion of the work’s contribution to knowledge and opportunities for future work. The chapter is structured as follows:

7.1 Synthesis of Research: This section discusses the various activities that contributed to the design, development and evaluation of the research presented in this thesis. The various models, constructs and biases applied in this research are discussed in terms of their contribution to the research.

7.2 Limitations and Considerations: The limitations of this research are discussed, including a consideration of the association with the eCute project and the issues of replicating this research. The role of the evaluator and the limitations imposed by the evaluand and evaluation context are considered. Reflections on possible improvements to the research design are discussed.

7.3 Originality and contribution to knowledge: The response to answering the research question is further considered, identifying the originality of this research. A discussion of the contribution to knowledge across multiple domains and the implications that arise is provided.

7.4 Future Work: The section discusses the advancement of the research presented in this thesis, focusing on further improving evaluation, for example through extending the evaluation approach, incorporating gamification and embedding evaluation activities within the application being evaluated. Further work on the use of scales to evaluate children is also considered.

7.5 Reflection: This section reflects on the completed research, providing views of both personal and professional development as a result of having completed the PhD.

7.1 Synthesis of Research

This section considers the various activities that contributed to the design, development and evaluation of the research presented in this thesis, discussing the impact in terms of addressing the fundamental question of this research, *“Are the outputs of evaluation with questionnaires improved when participant engagement informs questionnaire design?”*

7.1.1 Contribution from Literature

From the literature review a set of models, constructs and biases were selected for their relevance to this research:

Satisficing Model

The three elements, (task difficulty, ability and motivation) as shown in Krosnick's (1991) formula (figure 7.1) were used as a guide to reducing satisficing:

$$P(\text{Satisficing}) = \frac{a_1(\text{Task Difficulty})}{a_2(\text{Ability}) \times a_3(\text{Motivation})}$$

Equation 7.1: Krosnick's (1991), formula of satisficing.

Through a reverse engineering approach to Krosnick's formula it was identified that in order to reduce satisficing, task difficulty must be low, tasks must match the ability of the participants and motivation should be high, as indicated on the annotated version of Krosnick's formula in figure 7.2:

The diagram shows the formula $P(\text{Satisficing}) = \frac{a_1(\text{Task Difficulty})}{a_2(\text{Ability}) \times a_3(\text{Motivation})}$ with three annotations:

- An orange arrow points to $a_1(\text{Task Difficulty})$ with the text: "Task difficulty reduced, Language improved in Language study".
- A blue arrow points to $a_2(\text{Ability})$ with the text: "Evaluation materials designed to match ability age range in language and aesthetic".
- A green arrow points to $a_3(\text{Motivation})$ with the text: "Attempt made to address motivation by creating engaging experiences".

Figure 7.2: Annotated version of Krosnick's Model

The language used in the questionnaires was improved (see section 4.4) reducing the difficulty of the task by providing language that matched participant age range and anticipated ability levels. The questionnaire design workshop and review of children's media (appendix H) aimed to ensure that motivation was addressed by creating fun, novel and aesthetically appealing evaluation materials that engaged participants. Motivation is particularly challenging with children as a user group, with motivations such as contribution to the extension of knowledge or to the general good having little relevance for children (Chandler & Connell 1987). However, as the studies and engagement with the workbooks reveals, it is possible to motivate children to participate in evaluation because it is fun, engaging and interesting.

Optimal Response Model

The four phase optimal response model (Krosnick 1991; Bell 2007) is shown in figure 7.3.



Tourangeau & Rasinski (1988), 4 stages of question answering

This model was used in this research to underpin question answering. This model was selected as the work of Tourangeau & Rasinski, (1988), is a

seminal work in the field of evaluation. Phase one of the optimal response model inspired the language study (section 4.4) as a way of improving the familiarity of words used in the questionnaires. The second and third phases of the model required a method of stopping/slowing down the children to encourage them to consider their response before communicating it. Stickers were used as a method to enable communication of the judgement response in a novel way. Stickers not only slowed children before responding but children were also seen to pause and seemed to consider their response more deeply. Indicating that optimising had occurred and an optimal response had been given. These approaches clearly engaged the children and the variance and completeness of their responses highlights how engagement can impact upon response, as visualised in the annotated model of optimal response as shown below.

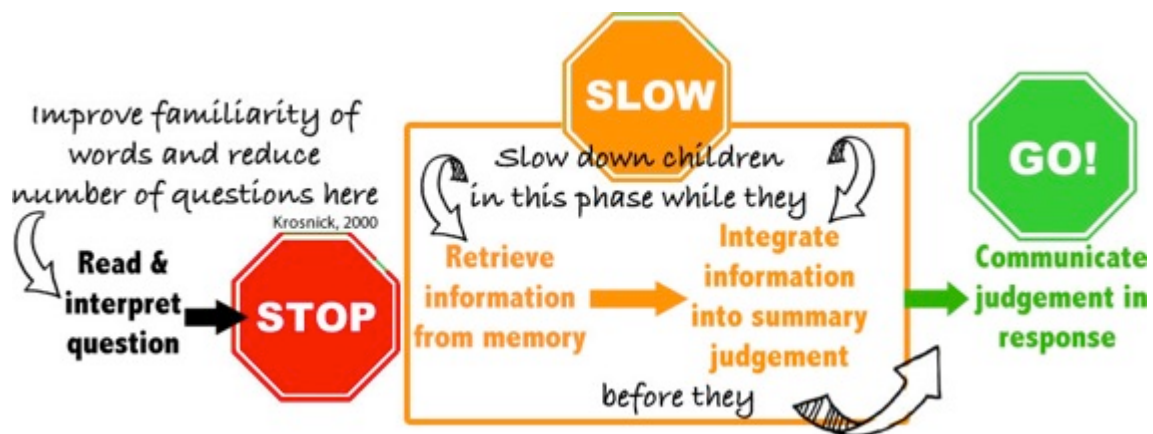


Figure 7.4: Annotated version of 'optimised' four stages of question answering

Model of Engagement

This research partially applied Attfield, Kazai, Lalmas, & Piwowarski's, (2011) model of engagement tailoring it to the evaluation context by incorporating focused attention; positive affect; aesthetics; durability/returnance and novelty (see section 2.5.2)

The positive affect, aesthetic appeal and returnance of the evaluation, were assessed through the postcard study (see section 6.2). The design of the postcard that collected the children's opinions of the workbooks had to tread lightly in terms of the amount of data collected and the complexity of the approach of assessing the children's engagement. Particular consideration was given to not overburdening the children with evaluation tasks, (Hanna & Ridsen 1997) hence a short and snappy postcard was used.

Following the approach to optimal response, the familiarity of words was improved, with the constructs simplified in the postcard questions to match users linguistic characteristics. Positive effect was referred to as fun, i.e. 'Was the workbook fun to do?' Aesthetic appeal by asking if 'The workbooks looked good?' Returnance was addressed by asking if the child would like to do another workbook. These simplistic measures were appropriate for the aim of the research, to evaluate aesthetic appeal, fun and returnance, and appropriate for the ability of the child participants.

Novelty was applied in the design of the workbooks. Although the workbook concept itself is not novel for children (see media review in appendix H), the

design provided novelty in experience, with variety amongst activities, page aesthetic and response formats.

Focussed attention was used as a measure of engagement in the evaluation. It was hypothesised that if the children were engaged then their attention would be focused. Focussed attention was assessed through the completion rates of the workbooks. As detailed in tables 7.5 and 7.6 the completion rate was high. In workbook one the CQS achieved 99.3% completion, the Messy was 100% complete and Bryant ranged from 97.81% to 100% complete. In workbook three the CQS was 100% complete, Messy 98.44% - 100% complete and Bryant 98.44% - 100%. With very few questions missed/skipped and all pages at least partially completed, contributing to the aim of collecting high quality data.

Response biases

Although a wide range of response biases exist, four were investigated in this research, with these four selected as being the most relevant to the research context, that being evaluation with children by questionnaire. The four response biases that this research sought to address were:

- **Social desirability:** participants may not accurately respond to questions as they are aiming to have socially desirable characteristics in order to appear more appealing to researchers (Oerke & Bogner 2011)
- **Acquiescence bias:** a tendency of respondent's to agree or respond positively (Danner et al. 2015) in the evaluation of self, products or services.

- **Satisficing:** respondents decide on and carry out (either consciously or unconsciously) a course of action that will satisfy the minimum requirements necessary to achieve a particular task (Horton, 2013).
- **Straight lining:** respondents provide responses at the same, usually extreme, point throughout the scale to either agree or disagree with the statements provided (Babbitt 1989).

From analysis of questionnaire responses, children did not feel the effects of social desirability or acquiescence bias, with most children exhibiting individual variance, and sample variance as a whole. This indicates that the effects of social desirability bias were low and children felt free to respond truthfully using the improved likert scale.

The remaining two response biases, satisficing, and straight lining, were reduced significantly (with results showing almost no straight lining) through the design of the workbook pages/activities. The approach to measuring the impact of this design was assessed through response variance in the workbook data. It was hypothesised that if the children had satisficed or acquiesced then the bias of straight lining (particularly towards positive ratings) would increase. With results highlighting variance in both individual and sample responses, little straight lining occurred.

The workbooks showed considerable variance. In addition, the majority of evaluation questionnaires were close to 100% complete. The children readily engaged with the workbooks as seen through observation, discussed in the CDFs and in data quality assessment. This was a result of a range of factors: the workbook appearance i.e. aesthetically appropriate, appealing and

engaging; the layout i.e. straight lines were limited by placing the scales on curved lines etc. and a redesigned Likert scale encouraged children to give a more varied and honest response. The evidence implies that the variance was due to a combination of all of these elements that contributed to the user centred design of evaluation.

7.1.2 User Engagement in Evaluation Design

In chapters 4 and 5, evaluation design studies were undertaken to inform and refine the evaluation materials. Table 8.1 details the 5 preliminary studies engaging the users in the design of evaluation materials.

Study	Focus
Questionnaire Design Workshop	To understand what is or is not engaging about standard questionnaires from the perspective of a 9 to 11 year old child. A discussion forum and a user centred design study of questionnaires.
Five Degrees- Visual Likert Scale Development	A series of studies to develop a visual Likert scale
Language Study	A study/exercise to test and improve the language used in the 3 questionnaires selected by eCute.
Engaging with questionnaires: Stickers as a response method	Investigating the use of stickers as an alternative to pen/pencil as a method of answering questions
Visual questions & answers: Nine Square	An investigation of an alternative method of collecting qualitative data.

Table 7.1: Preliminary Studies

The research presented in this thesis and publications (e.g. Hall, Hume, & Tazzyman, 2014, Hall & Hume, 2012) provides one of the only examples of users being engaged in the design of evaluation.

The questionnaire focus group confirmed the theory/hypothesis of this research, that children are disengaged by standard format questionnaire documents. In response, the final evaluation questionnaires and experimental protocol were designed to reduce the sense that children were participating as subjects in an experimental study or that they were filling in questionnaires.

Early stage prototypes during workbook development used the traditional questionnaire grid format and although aesthetically different to most questionnaires they reflected earlier work on improving questionnaire design by superficial improvements with ORIENT (Hall et al., 2013). Instead, the research in this thesis, improves engagement not only through aesthetics, but further through user-centred design targeting straight lining, satisficing and sub-optimal responses. For example, during the questionnaire focus group the children expressed that they enjoy the Likert scale response format, a view supported in other research (van Laerhoven et al. 2004; Haddad et al. 2012; Mellor & Moore 2014). However, one child reported that they enjoyed going “tick, tick, tick...” when responding, providing a direct example suggesting that a fully optimised response was not given and that straight lining had occurred. The straight lining problem was resolved by removing linearity from the questionnaire layouts, with questions placed on curved and wiggly lines and also staggered across pages.

Unsurprisingly, children felt strongly about adding colour and decoration to the questionnaires, with the questionnaire design workshop revealing what children expect and desire from a questionnaire in terms of visual design. The

design direction taken by the children was directly reflected in the final design of the workbooks. In the questionnaire workshop many children produced booklets, rather than individual pages; therefore the final evaluation materials were designed in a similar way as workbooks. The children's designs were full colour, with a lot of decoration and images added to match the subject theme that they had chosen. This was then mirrored in the design of the workbooks with graphics added to match the narrative of the MIXER application. The review of children's media also highlighted the need for the vibrant materials, and as can be seen the evaluation materials have a similar aesthetic appeal to many hard copy activity comics / magazines targeting children. This, however, is not novel for the children, it is simply an expectation; children expect content targeted at them to look appealing and appropriate.

Children in the questionnaire focus group, workshop and pilots of the questionnaires viewed the Likert scale positively. The scale designed in the Five Degrees study successfully elicited a full range of responses from the children. The sequentially iterative approach taken, using different groups of children, was effective, with the final scale from the Five Degrees study used in both the workbooks and the evaluation postcard. The scale successfully gathered data that showed individual and sample variance in responses with children using all five points of the scale to respond.

A number of techniques and approaches were used to engage through fun, including the use of quizzes, following the route of the questions and stickers. The 'Which woodland animal are you?' (CQS, workbook one) received a

strong positive reaction, with children very eager to find out which animal group they were in and carrying on the role of a badger, deer or fox during the break.

The children's reaction to the stickers was surprising; the children were a lot more enthused by receiving the stickers than had been expected. It may be that while teachers frequently use stickers as a reward system to acknowledge good work or behaviour, they are not something that children themselves often get to work with.



Table 7.2: Children applying stickers to the workbooks

Another finding relating to the use of stickers was the observation of the children pausing before a) selecting a sticker and b) before finally affixing the sticker to the page. In terms of children providing an optimal response (Krosnick 1991; Bell 2007) the 'pause' aligned with the 'slow down' phase of the annotated model of optimal response. Rather than jump from 'read & interpret question' directly to 'communicate judgement response' or at worst jump directly to 'communicate judgement response'. The children were observed to read the question, *pause*, contemplate which sticker to select, and *pause* again before fixing it onto the page.

The Nine Square study investigated an alternative approach to qualitative data collection, the study used large format poster sized comic strips, the study was replicated on a smaller individual scale in the workbooks in an activity called The Trip. Children provided strong and clear narratives, with detailed drawings reflecting the children's feeling about going on trips and meeting new people.

The final workbooks, with example pages shown in figure 7.2 were extremely well received by children, teachers, stakeholders and the R&D team. Children's attachment to the workbooks was surprising. Pilots of the workbooks highlighted that children did not want to hand over the workbooks at the end of the session. To partially meet this expectation and to give the children a "take away" the last page of the workbook that contained the word search was removed (as it contained no data) and given to children so they could complete it later.

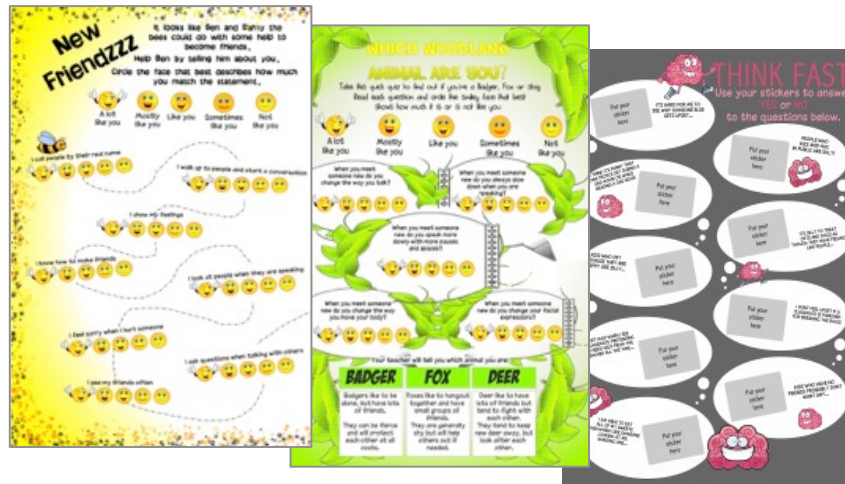


Figure 7.1: Example pages from the final workbooks

7.1.3 Discussion of Main Findings

The aim of the evaluation was to provide an enjoyable and engaging experience for children, which gathered high quality data. The meta-evaluation consisted of three phases; 1) measures of completion rates and variance in responses provided in the workbooks, 2) a feedback style postcard and 3) a Classroom Discussion Forum (CDF). The key findings from the meta-evaluation were:

Data quality:

- The completion rate of the workbooks was very high, ranging from 97.81% (one child did not complete 4 of the questions) to 100% complete.
- Variance was also high in the children’s responses, with the majority of children either using the entire scale or using both yes and no to respond in Bryant’s Empathy Index

Postcard findings:

- 74% (n = 87) of children rated the workbook as fun to do
- 65.6% (n=78) of children agreed the workbook looked good
- 70.6% (n = 92) of children gave a positive indication that they would like to repeat their evaluation experience
- 38.6% (n = 46) of children selected an evaluation activity as their favourite with The Trip activity selected the most, followed by the 'Which woodland animal are you?' activity.

CDF Findings:

- Children stated that they enjoyed the workbooks. They also expressed that they enjoyed the stickers very much and the "Which woodland animal are you" activity.
- Children expressed a desire to repeat the evaluation experience and to do another workbook

The data quality suggests that children were engaged with the workbooks, as completion rates were high. The variance in responses indicates that optimal responses were given and that the majority of children did not succumb to response biases such as satisficing, straight lining, social desirability or acquiescence bias. The children's self report data in which the majority of children state that the workbooks were fun and were an experience that they wish to repeat corroborate the engagement exhibited in the high quality of the data. Children ratings of the aesthetic appeal, (looked good?), of the workbooks was lower than anticipated, however, during the CDF the children stated that the workbooks were very similar to their recreational literature. The interpretation of this finding was that aesthetic design of the workbooks *met* rather than *exceeded* the children's expectation.

7.2 Limitations & Considerations

7.2.1 eCute

As detailed in appendix A this research used the MIXER application (from the eCute project) as the evaluand and evaluated MIXER in line with the projects' evaluation requirements. Whilst the involvement with the eCute project was mutually beneficial, the relationship did bring with it limitations in the form of a set of stipulations from the eCute R&D team that had to be adhered to. These were:

Questionnaires, Learning goals and research design - The questionnaires selected by eCute were the CQS (Ang, et al., 2007), Bryant's Empathy Index (Bryant 1982), the MESSY Scale (Matson et al. 1983) and the EEQ (L. Hall et al. 2015) all are well known, frequently used and validated questionnaires. The questionnaires were considered and selected by experts in the field of psychology and education, and were a positive inclusion to this research.

Classroom context and children as participant group - The participant group used in this research were children aged 9-11. The limitations that arose from the use of this participant group were absences of children due to illness. The children were very mixed in terms of ability, with some children finishing the workbooks quickly with no help and others needing help to read the questions. As the studies were conducted in schools problems arose from the staff not

fully understanding the experimental conditions and the importance all children being tested under same conditions. In one class a teacher allowed a child to leave the classroom for music group, therefore this child's data was incomplete.

Hard copy materials - There was also a stipulation that the evaluation materials were hard copy format only. At times this added pressure to the delivery of the evaluation sessions with the printing and collating of the workbooks requiring considerable time and effort. A digital format would have reduced this effort and also reduced costs in terms of ink and paper etc.

While these stipulations were limiting they were seen as positive and authentic constraint upon this research. Professional evaluators (in industry) would be called upon to evaluate applications or programs with similar stipulations; this gives the research authenticity and validity outside of the research context. Other limitations that resulted from the involvement and dependency on the R&D team included delays in the delivery of the MIXER application due to technical and development problems.

Developing and research evaluation materials to the extent taken in this research was far beyond the scope of the research aims of the eCute project. The distinction between the work related to this research and the work of the eCute project are summarised in figure 7.3.

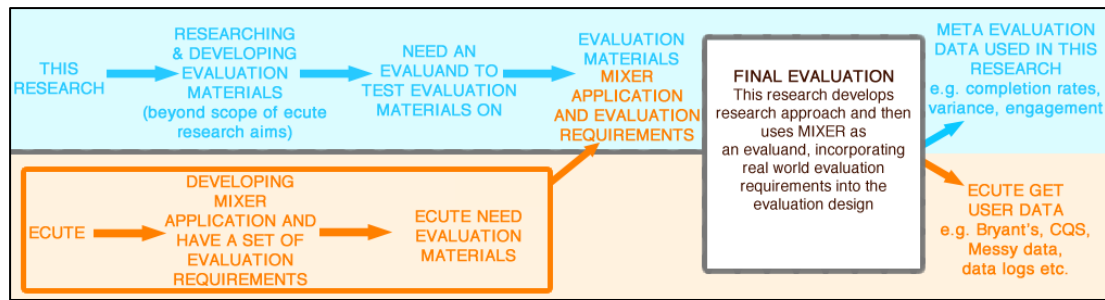


Figure 7.2: Distinction between this research and the eCute Project

7.2.2 Methodological Limitations and Considerations

A potential limitation of this work relates to the focus on children. This raises the question of whether this research is relevant and generalizable to all user populations. Although response biases, aesthetic design and user expectations for adults may be different to those seen with children, the basic principle of user-centric evaluation, (that of designing evaluation instruments for users as well as for the R&D team), holds. Providing a user-centric evaluation experience would undoubtedly improve the experience for the user and would increase their engagement in the evaluation process.

A potential limitation of this research in terms of its replicability is the amount of time and work involved in producing the evaluation materials. In addition to the review of academic literature and children’s media and the preliminary studies that contributed to the final workbooks, there were many iterations required. The workbooks were incrementally improved to reduce straight lining or improve the age appropriateness and appeal of the aesthetic etc., some early examples are shown in (figure 8.6). However, this time and effort

clearly impacts on the quality of the data produced and the children's engagement.

The workbooks were designed using a Participatory Design approach, with components being tested and incrementally improved over time as in the development of an interactive application. In a research project where the primary aim is to develop an application there may not be resource available to dedicate to such detailed consideration of the evaluation materials, or a graphic designer may not be part of the R&D team and this may require costly outsourcing of the design of the materials. This was the case in the eCute Project, the research and development of evaluation materials as developed in this research was outside the scope of the research aims of the project.

However, this research puts forward the argument that the evaluation of any application being developed should be given equal consideration in terms of participant engagement and interaction approach. This argument for equal consideration is justified by the results of this research as reported in chapter 6, that show children were engaged resulting in reliable, quality data that was almost 100% complete and that showed sample and individual variance indicating optimal responses had been provided.

As a test of influence on social desirability bias the children should have been asked if they had felt that they were taking part in an experiment. An additional evaluation element that would have improved this research would have been to repeat the feedback postcard that followed workbook one. This evaluation activity should have been repeated after each workbook to gather

qualitative and quantitative data that would have indicated if the children's engagement increased, stayed the same or reduced after workbooks two and three. While the CDF that followed workbook three included the same themes the responses were positive, and from observing the children during completion of workbook two and three they were very focused during completion and excited when talking about the workbooks during the CDF, indicating sustained engagement with the workbooks. However, the additional quantitative data would prove a stronger case for the children's engagement with the workbooks.

Another addition that would have improved the data collected would have been to provide a multiple-choice activity, listing all of the names of the activities from the workbooks. This would have reduced any uncertainty in responses such as the two children who gave the response 'smiley faces' as their favourite activity. This approach could be further refined to only offer choices of evaluation activities. Over 50% of children responded that the word search was their favourite activity; on reflection this choice could have been anticipated as word search activities are known to be popular with children.

7.2.3 Evaluator Role

While the design of the workbooks and activities contained within them were very popular, consideration must also be given to the impact of the evaluation team delivering the evaluation session.

This aspect of the evaluation brings into play more than the application of a set of design principles and constructs that aim to induce engagement. Evaluators in this context require knowledge of evaluation and of the subject domain of the evaluation. The ability to design and follow an experimental protocol (without it appearing as so to the children) is also required. Evaluators must be natural and calm around groups of 30+ very excitable children. Evaluators must also be able to communicate with and control the classroom context whilst keeping children's attention focussed and simultaneously resetting a crashed application etc. The evaluator in this research had all of the skills listed and more, bringing together a combination of experiences gained over many combined years of previous research experience, teaching, working with and raising children. Any evaluator or evaluation team form a variable in research that may be influential in the positive results reported in this research which and may prove difficult to replicate in another study.

Some aspects of designing an engaging questionnaire are similar to the design considerations given when designing an interactive application, for example, an age appropriate aesthetic and consideration of layout. However, there are also aspects of the design approach that are very different. In designing an interactive application the options are far greater, with novelty provided through sound, movement and illustration etc. being employed to surprise, amuse and engage the user, these are obviously not applicable to hard copy materials.

7.3 Originality and Contribution to Knowledge

This thesis aimed to answer the research question: *“Are the outputs of evaluation with questionnaires improved when participant engagement informs questionnaire design?”*

As this research has identified, evaluation can be engaging, even with the limitations of a hard copy medium, evaluation can create an enjoyable, engaging experience rather than a disruptive, disappointing questionnaire completion session.

This research has taken the same constructs and consideration of engagement applied in the development of an application and has applied them to the design and development of an evaluation. The hypothesis underpinning this research was: *The outputs of evaluation will be improved when participant engagement informs questionnaire design.* This research has answered the hypothesis by proving it to be true. Clearly, *the outputs of evaluation are improved when participant engagement informs questionnaire design.*

The main originality and contribution to knowledge of this thesis are summarised here and further discussed below.

- Placing the user at the centre of the evaluation process though applying UCD, UX and HCI techniques
- Focusing on user engagement in the evaluation experience and explicitly designing in engagement to design out response biases
- Approaches, informed by literature and empirical studies, to gain

quality data from children using quantitative survey instruments

- Transdisciplinary contribution, with approaches, findings and results relevant to experience focused evaluations and the evaluation of those evaluations.

Although many interactive systems are evaluated, there has been little previous research relating to involving users in the design of the evaluation experience. Whilst the role and approaches of evaluators have been well considered, for example Heuristic Evaluation (Cockton & Woolrych 2001; Woolrych et al. 2003) and user testing (Hertzum et al. 2014; Nielsen 1994), the perspective, requirements and expectations of the user have received little interest. Whilst it is well recognised that users need to be at the centre of the interaction design process, their role in evaluation has been (and typically continues to be) the subject in an experiment. This research has challenged this relegation of user from the centre to periphery and is one of the first to have considered and applied HCI methods to the design of evaluation.

This research is original in considering engagement in evaluation as a method of designing out and reducing response bias, synthesising and contributing knowledge to evaluation. This research identifies that when engagement informs evaluation design and when this is combined with a design response to reduce response biases that an increase of engagement and improvement of data quality will follow. This is seen through observation, with children highly engaged in both the preliminary studies and quantitatively and empirically derived, with workbook data highlighting significant variance in responses. This demonstrates a reduction in responses biases and the

provision of quality data from engaged participants demonstrated through results such as an almost 100% completion rate from children.



Figure 7.3: Model of evaluation informed by engagement

This approach challenges current methods of evaluating with children. Instead, it adopts an innovative approach, using the methods and techniques from Participatory Design applied to the design of evaluations for children rather than for the data users. The iterative method of user requirements > design and develop with users > test with users > iterate > test etc. using methods and techniques informed by literature and knowledge relating to user characteristics and biases produces engaging evaluations.

Perhaps, the most significant implications and potential contributions arising from this research are the approaches to designing out the response biases of satisficing, acquiescence and social desirability. These approaches are easy to replicate for the majority of quantitative survey instruments targeted at children. Of particular consideration for both originality and contribution has been this thesis' consideration of children's responses to Likert scales (Hall et al. 2016). For example, this research identified that children respond with

greater variance when negative faces are removed, with the Five Degrees scale offering a novel contribution to evaluators.

Through addressing the response biases of straight lining, acquiescence and social desirability, optimal answering of questions was increased. It is well known that children have response biases is well known (Read et al. 2002a; Bell 2007), yet there has been little prior work to resolve them. Through resolving these biases, this research challenges many of the results published about interactive narrative based systems (and others) evaluated by children.

It particularly casts doubt on the validity of results obtained with linear/grid-format questionnaires and rating scales. It can be suggested that the positive indications given in many child evaluation studies, showing preference or agreement etc., could be somewhat less positive than interpreted. With an increasing use of quantitative instruments to gain children's data (Greig et al. 2012) the Five Degrees scale and approaches to avoid linear and grid formats provide an important contribution to the evaluation community.

Through the dissemination of the research in this thesis, there has been a contribution to current evaluation practises with children. For example, through the association with the eCute project the workbooks have been used by to successfully evaluate MIXER with children in Germany, Australia, Portugal and Japan. Further, the EU rated eCute as Excellent in its final review, with the MIXER evaluation instruments and approach highlighted as excellent practice. Publications that have emerged from this thesis are provided in Appendix L and include:

- Hall, L., Hume, C. and Tazzyman, S., 2016, June. Five Degrees of Happiness: Effective Smiley Face Likert Scales for Evaluating with Children. In *Proceedings of the The 15th International Conference on Interaction Design and Children* (pp.311-321). ACM.
- Hall, L., Hume, C. and Tazzyman, S., 2015. Engaging Children in Interactive Application Evaluation. *Enfance*, 2015(01), (pp.35-66).
- Endrass, B., Hall, L., Hume, C., Tazzyman, S. and André, E., 2014, June. A Pictorial Interaction Language for Children to Communicate with Cultural Virtual Characters. In *International Conference on Human-Computer Interaction* (pp.532-543). Springer International Publishing.
- Hall, M., Hall, L., Hodgson, J., Hume, C. and Humphries, L., 2012, April. Scaffolding the Story Creation Process. In *CSEDU (1)* (pp.229-234).
- Hall, L. and Hume, C., 2011, October. Why numbers, invites and visits are not enough: Evaluating the user experience in social eco-systems. In *SOTICS 2011, the first international conference on social eco-informatics* (pp.8-13).

This research has resulted in evaluation materials and experiences that engage children. Achieving engaging materials is feasible and achievable, yet requires an effort that is currently not dedicated to evaluation design in interactive system evaluation. And more, it requires a different perspective of evaluation, one where the subject matters as much as the data they produce.

This research has aimed to design evaluations for children as well as for the data users. This research included the evaluation of an evaluation considered from both a user and a data user perspective. This provides an easily applicable mixed methods approach that considers both the user experience

and the data quality. Using the postcard measure to gain both qualitative and quantitative user responses to the evaluation experience complemented by observation and focus groups, children's engagement was clearly established. Through applying completeness, individual and sample variance across the workbook data, optimal responding could be seen.

This research provides a transdisciplinary contribution to knowledge. It clearly identifies that evaluation is a transdiscipline, with the research informed from and relevant to fields including media, computing and psychology. As such this research has not only contributed to the domain of evaluation, but also to the domains of HCI, education and psychology, and to a broader extent any domain that conducts evaluation with children.

7.4 Future Work

There are many ways in which this research could be extended, with a myriad of areas offering future directions across a range of disciplines. Here, the focus is on the near future of this evaluation research.

An area of particular interest would be to explore the impact of medium on user response to evaluation. This could include the provision of the materials in a digital format, comparing the impact of screen to paper. For example, the workbooks could be turned into an interactive application with animation and sound being added to further increase engagement with the evaluation. This would most likely result in new biases and engagement inhibitors requiring

further exploration as to how to design these out and to enable optimal responses.

Alternative directions could also involve embedding the evaluation directly within the narrative and/or the interaction experience. For example, with MIXER, by having Tom ask the evaluation questions and the child responding via the iPad. Although intuitively this seems sensible, it could quickly become irritating to the user. Thus, creating verbal questionnaires to replicate instruments such as the CQS would require considerably more consideration than a simple change from paper to Tom's verbalisation. This approach would also require a possible reconsideration of the evaluation of the evaluation, with very different questions and issues emerging. For example does increasing interaction with Tom to support evaluation impact on the child's perspective of Tom? This could be perhaps negative, with Tom's limitations such as synthetic speech reducing his believability and likeability; or positively with the child increasing their empathic response or attributing greater capacity to Tom.

During the review of engagement, gamification was identified as an approach to increasing and maintaining user engagement. Gamification is currently a very popular research topic and a significant amount of time was spent reviewing gamification approaches and applications and their relevance to this research. While gamification was indirectly influential some inspiration was taken from the gamification approach, this can be seen in the YES or NO activity (section 6.3.5) which slightly resembles a board game layout, the

initial use of stickers was inspired by badges (a common feature in gamification) as a reward system well recognised by children. The gamification of evaluation could be one possible direction for future work; for example, awarding points or badges for every evaluation activity completed could increase engagement and motivation. This would prove useful in pre-, in and post-test evaluation studies where repeat evaluation is required.

As discussed in limitations, this research was conducted with children aged 9-11. Future work could consider the designing out of response biases in quantitative survey instruments across children of younger and older age ranges and with adults. It would be useful to identify and explore age-dependant biases and to illustrate how these could be resolved in appropriately designed evaluations.

The user centred evaluation approach promoted in this thesis, applying UCD techniques and approaches to evaluation rather than interaction, has clear resonance for the design of any user experience evaluation. Applying this approach to a range of user groups interacting with a diversity of narrative systems would enable the generalizability of this user centred evaluation approach to be explored.

As discussed in section 7.3 the Likert scale developed in this research has a significant implication for previous research findings with children. Further research is planned on the Five Degrees scale, aiming to investigate its applicability to other evaluation areas, such as education, children's product views and health. This is related to research currently being developed in

collaboration with psychologists, aiming to validate the Five Degrees scale from an R&D perspective. This will aim to ensure that researchers accept and use the scale by providing the necessary psychological validation and reliability assessments within a traditional publication targeting the research community.

An area of future research is to explore the data quality of existing datasets where Likert scales have been used for child user experience evaluations. Through reviewing completeness and variance of datasets the destructive impact of biases on data quality may be identifiable and the issue of data quality further investigated. The approach to evaluating evaluations presented in this thesis is currently being disseminated and promoted, with the aim of engaging with other R&D teams to evaluate their approaches.

7.5 Reflection

In terms of the subject matter presented in this thesis I now feel very differently about evaluation. When I began this research I viewed evaluation as giving someone a set of questions, forming a questionnaire, to generate results for the R&D team. I now understand there is much more to evaluation and this thesis has investigated and presented a mere fraction of the research required to bring evaluation the focus I believe it deserves as a scientific discipline in its own right.

Even acknowledging the extensive research that is still needed, I can clearly identify that this research made a real difference to the children participating in the MIXER evaluation. Although the findings are not included in this thesis, I was present at the German evaluation of the MIXER application and I was pleased to observe that the engagement response with the evaluation materials was similar to the engagement of children in the UK. Children focussed on completing the workbooks and reacting with great joy at the use of the stickers. At break time the children in Germany began trading the remaining stickers to collect a full set of characters, this suggests engagement with the MIXER application and the evaluation materials. This experience highlights that it is not just the evaluators who made the evaluation engaging which was one of my concerns discussed in limitations, but rather that it is the evaluation itself that is engaging. Thus, the originality and contribution of my work, taking evaluation from a session disliked by users to an engaging, complementary activity that users wanted to repeat, is significant and offers a new perspective of the role and nature of the evaluation of interactive, narrative-based systems.

7.6 Summary

This chapter discussed the various activities that contributed to the design, development and evaluation of the research presented in this thesis and how their contribution shaped this research. Limitations, including the impact of the association with the eCute project, methodological considerations and

evaluator role were considered, along with potential improvements and areas for future work. The originality and contribution of this research was outlined, clearly highlighting the potential of the approach presented in this thesis to increase user engagement in evaluation. The following chapter will conclude this thesis.

8 CONCLUSIONS

This thesis sought to answer the research question, “*Are the outputs of evaluation with questionnaires improved when participant engagement informs questionnaire design?*” In doing so, an exploration of the impact of participant engagement in evaluation has been presented. In this final chapter, the main conclusions drawn from this research are presented.

Firstly, this research concludes that users should always be placed at the centre of the evaluation design just as they are often central to the design of an interactive system. In system design applying a user centred or participatory design approach ensures that the system designed reflects the desires, needs and ability of the intended end users. It seems obvious that the same approach should be applied in evaluation, with ability, motivation, needs and desires of participants considered in the design of evaluation, yet this approach is not reported in literature.

Secondly, this research concludes that designing evaluation experiences that are informed by participant engagement i.e. providing evaluation materials that are aesthetically appealing, fun, novel and age appropriate in content and in context, can reduce and perhaps remove response bias. Satisficing, acquiescence bias and straight lining can result in reduced data quality, yet as this research has demonstrated, even simple design changes such as removing linearity from layouts can reduce occurrences of straight lining. By improving the language used to match the ability of respondents’ satisficing is reduced. Social desirability and acquiescence are removed by providing an

evaluation context that is natural for children by designing for participants in the role of a school child. Additionally, through the redesigned Smiley Face Likert scale, children were free to respond openly and honestly to questions.

Thirdly, this research concludes that evaluations both need to be, and can be, improved for children. Greater consideration of user characteristics, expectations and experience in evaluation design significantly impacts upon the look and feel of evaluation instruments. This is of particular importance for children, for just as it would not be expected that children would engage with the same media, games and experiences as adults, so too, should their evaluations be tailored and designed to meet their characteristics, needs and expectations. This research has clearly outlined how such characteristics, needs and expectations can be incorporated through developing evaluations that aim to engage children as well as to inform researchers.

Fourthly, this research concludes that increasing engagement in evaluation does have a positive impact on data quality. In this research data quality was measured in terms of completeness of data provided and variance within that data both at the individual and sample levels. By considering the layout of questionnaires, for example, by designing questionnaire layouts that stagger questions across the page or provide lines to follow from one question to the next, fewer questions are missed which contributes to full and complete data sets. Using stickers contributed to the provision of optimal responses by slowing down participants between reading and responding to questions. The

quality of the data collected in this research is evidenced with almost complete data sets that show variance in both individual and sample data.

This research evidences that when engagement informs evaluation that the experience and outcomes of evaluation are improved for all concerned. The experience is enjoyable for participants, this is evidenced in the postcard study in section 6.2, with 74% of children giving a positive indication that they had found the experience to be fun and 70.6% of children indicating that they would like to repeat their evaluation experience. The outcomes for the R&D team are improved with data quality that is almost 100% complete, with optimal responses that are more considered, showing individual and sample variance, indicating a reduction and at best complete removal of common response biases.

Finally, this research concludes that evaluation is under researched and under considered and more research is needed. In the review of evaluation it was evidenced that there has been little research or innovation in the domain of evaluation. Along with many others, when beginning this research I was of the opinion that evaluation was something that was separate to the interaction. I viewed evaluation as an add on that came at the end to answer research questions; to assess that the user had enjoyed themselves; and to evaluate that the application performed as required, with the application very much at the forefront at all times. Having completed this research I now think differently. Evaluation, and its design, should be considered to be as important as the design of the interaction with the evaluand, and requires

significantly more consideration with this largely ignored field offering considerable potential for future research.

Evaluation is a transdisciplinary domain and this research has taken existing methods and models from HCI, user experience, psychology and education and applied them in novel ways. The methods used have existed for many years, yet few researchers have considered aggregating and applying them to evaluation. This research concludes that evaluation can be improved for the user by taking a user-centric approach and designing in engagement both in terms of visual appeal and in responding to user characteristics through reducing response bias. The hypothesis upon which this research was conducted states that the outputs of evaluation will be improved when participant engagement informs questionnaire design. The research presented in this thesis has proven this hypothesis with an evidenced, novel and transdisciplinary contribution to the domains of HCI, education, psychology and social sciences, but most importantly it has also made a rare, valuable and much needed contribution to the domain of evaluation.

This research concludes that by designing engagement into evaluation two distinct and equally important goals can be achieved. Not only can R&D questions be answered with high data quality but further, this approach ensures that the users who help us to answer our research questions have an enjoyable and engaging experience.

References

- Ackroyd, S., 1992. *Data collection in context*, Longman Group United Kingdom.
- Aguirre, A. et al., 2014. Proposal to evaluate the satisfaction of use in Virtual Learning Environments. In *Proceedings of the XV International Conference on Human Computer Interaction*. ACM. p. 28.
- Aldridge, S. & Rowley, J., 1998. Measuring customer satisfaction in higher education. *Quality Assurance in Education*, 6(4), pp.197–204.
- Alkhafaji, S. & Sriram, B., 2012. Educational Software Development Life Cycle Stages. *Chinese Business Review*, 11, pp.128–137.
- Ang, S. et al., 2007. Cultural Intelligence: Its Measurement and Effects on Cultural Judgment and Decision Making, Cultural Adaptation and Task Performance. *Management and Organization Review*, 3(3), pp.335–371.
- Anić, I., 2015. Participatory Design: What is it, and what makes it so great? *UX Passion*. Available at: <http://www.uxpassion.com/blog/participatory-design-what-makes-it-great/> [Accessed January 3, 2016].
- Attfield, S., Piwowarski, B., Kazai, G., et al., 2011. Towards a science of user engagement (Position Paper). *Evaluation*.
- Autodesk, 2016. Designing the User Experience at Autodesk. *ww.dux.typepad.com*. Available at: http://dux.typepad.com/digital_library/methods/ [Accessed October 4, 2016].
- Aylett, R. et al., 2014. Werewolves, Cheats and Cultural Sensitivity. In *2014 International Conference on Autonomous Agents and Multiagent Systems*. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1085–1092.
- Babbitt, B., 1989. Questionnaire construction manual annex. Questionnaires: Literature survey and bibliography. *Operations Research Associates*.
- Bargas-Avila, J.A. & Hornbæk, K., 2011. Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. pp. 2689–2698.
- Barkhuus, L. & Rode, J.A., 2007. From Mice to Men—24 years of Evaluation in CHI. *CHI 2007*, pp.1–16..
- Bartle, R., 2004. *Designing Virtual Worlds* R. A. Reiser & J. V Dempsey, eds., New Riders Games.

- Behrendt, M. & Machtmes, K., 2016. Photovoice as an evaluation tool for student learning on a field trip. *Research in Science & Technological Education*, 34(2), pp.187–203.
- Bell, A., 2007. Designing and testing questionnaires for children. *Journal of Research in Nursing*, 12(5), pp.461–469.
- Bevan, N., 2012. Usability Evaluation Methods |. *Usability Body of Knowledge*.
- Bevan, N., 2009. What is the difference between the purpose of usability and user experience evaluation methods. In *Proceedings of the Workshop UXEM* (Vol. 9, pp. 1-4).
- Bian, Y. et al., 2016. A framework for physiological indicators of flow in VR games: construction and preliminary evaluation. *Personal and Ubiquitous Computing*, pp.1–12.
- Birkett, A., 2015. Survey Design 101: Choosing Survey Response Scales. *ConversionXL.com*.
- Black, G., 2005. *The Engaging Museum. Developing Museums for Visitor Involvement*, London, New York: Routledge Psychology Press.
- Bødker, S. et al., 2016. Advances in Participatory Design. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. pp. 984–987.
- Bossen, C., Dindler, C. & Iversen, O.S., 2016. Evaluation in Participatory Design: A Literature Survey. In *Proceedings of Participatory Design Conference*.
- Bryant, B., 1982a. An index of empathy for children and adolescents. *Child development*, 53(2), pp.413–425.
- Bryant, B., 1982b. Index of Empathy for Children and Adolescents Bryant, B. (1982). An Index of Empathy for Children and Adolescents,. *Children*.
- Bryman, A. & Bell, E., 2015. *Business research methods*, Oxford University Press, USA.
- Cairns, P., Cox, A. & Nordin, A.I., 2014. Immersion in digital games: a review of gaming experience research. *Handbook of digital games*, MC Angelides and H. Agius, Eds. Wiley-Blackwell, pp.339–361.
- Carey, M.A. & Asbury, J.-E., 2016. *Focus group research*, Routledge.
- Castro, J.R. & Gramzow, R.H., 2015. Rose colored webcam: Discrepancies in personality estimates and interview performance ratings. *Personality and Individual Differences*, 74, pp.202–207.

- Chandler, C.L. & Connell, J.P., 1987. Children's intrinsic, extrinsic and internalized motivation: A developmental study of children's reasons for liked and disliked behaviours. *British Journal of Developmental Psychology*, 5(4), pp.357–365.
- Chen, F. et al., 2016. Applications of Cognitive Load Measurement. In *Robust Multimodal Cognitive Load Measurement*. Springer, pp. 235–247.
- Chiewvanichakorn, R., Nossal, N. & Hiroyuli, L., 2015. Game refinement model with consideration on playing cost: A case study using crane games. In *7th International Conference on Knowledge and Smart Technology. IEEE*. pp. 87–92.
- Chisik, Y. & Mancini, C., 2016. Of kittens and kiddies: reflections on participatory design with small animals and small humans. In *Proceedings of the 14th Participatory Design Conference on Short Papers, Interactive Exhibitions, Workshops - PDC '16*. New York, New York, USA: ACM Press, pp. 123–124.
- Cockton, G. & Woolrych, A., 2001. Understanding inspection methods: lessons from an assessment of heuristic evaluation. In *in People and Computers*. pp. 171–192.
- Cole, J.S., McCormick, A.C. & Gonyea, R.M., 2012. Respondent use of straight-lining as a response strategy in education survey research: Prevalence and implications. In *Annual meeting of the American Educational Research Association*. pp. 1–18.
- Colombo, L. & Landoni, M., 2014. A Diary Study of Children's User Experience with eBooks Using Flow Theory as Framework. In *Proceedings of the 2014 conference on Interaction design and children - IDC '14*. pp. 135–144.
- Coombe, C. & Davidson, P., 2015. Constructing Questionnaires. *The Cambridge Guide to Research in Language Teaching and Learning*, p.217.
- Craig, S.D. et al., 2008. Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive–affective states during learning. *Cognition & Emotion*, 22(5), pp.777–788.
- Creswell, J.W., 2012. *Qualitative inquiry and research design: Choosing among five approaches*, Sage.
- D'Mello, S., Olney, A., Williams, C. and Hays, P., 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5), pp.377-398.
- Dahlgren, G.H. & Hansen, H., 2015. I'd rather be nice than honest: An experimental examination of social desirability bias in tourism surveys.

Journal of Vacation Marketing, 21(4), pp.318–325.

- Danner, D., Aichholzer, J. and Rammstedt, B., 2015. Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, 57, pp.119-130.
- DeSmet, A. et al., 2016. Is participatory design associated with the effectiveness of serious digital games for healthy lifestyle promotion? a meta-analysis. *Journal of medical Internet research*, 18(4).
- Dickinson, D.M. et al., 2016. The Language of Cigar Use: Focus Group Findings on Cigar Product Terminology. *Nicotine & Tobacco Research*, pp.285.
- Dirican, A.C. & Göktürk, M., 2011. Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, 3, pp.1361–1367.
- Douglas, Y. & Hargadon, A., 2000. The pleasure principle: immersion, engagement, flow. In *Proceedings of the eleventh ACM on Hypertext and hypermedia*. ACM, pp. 153–160.
- Drakou, M. & Lanitis, A., 2016. On the development and evaluation of a serious game for forensic examination training. In *2016 18th Mediterranean Electrotechnical Conference (MELECON)*. IEEE, pp. 1–6.
- Druin, A. et al., 1998. Children as Our Technology Design Partners. *The Design of Children's Technology*, (Age 8), pp.51–60. Available at: <http://drum.lib.umd.edu/handle/1903/947>.
- Endrass, B., Hall, L., Hume, C., Tazzyman, S. & Andre, E., 2014. A Pictorial Interaction Language for Children to Communicate with Cultural Virtual Characters. In *16th International Conference on Human Interaction*. Heraklion, Greece, p. in press.
- Endrass, B., Hall, L., Hume, C., Tazzyman, S., Andre, E., et al., 2014. Engaging with virtual characters using a pictorial interaction language. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. pp. 531–534.
- Fagerberg, P., Ståhl, A. & Höök, K., 2004. EMoto: Emotionally engaging interaction. *Personal and Ubiquitous Computing*, 8, pp.377–381.
- Ferreira, B.M. et al., 2016. Evaluation of UX Methods: Lessons Learned When Evaluating a Multi-user Mobile Application. In Springer International Publishing, pp. 279–290.
- Følstad, A., 2007. Group-based expert walkthrough. In *R3UEMs: Review, Report and Refine Usability Evaluation Methods. Proceedings of the 3rd. COST294-MAUSE International Workshop*. pp. 58–60.

- Frauenberger, C. et al., 2015. In pursuit of rigour and accountability in participatory design. *International Journal of Human Computer Studies*, 74, pp.93–106.
- Freeman, T., 2006. “Best practice” in focus group research: Making sense of different views. *Journal of Advanced Nursing*, 56(5), pp.491–497.
- Frølunde, L., 2014. 9 Reflexive Learning through Visual Methods. *Situated Design Methods*, p.161.
- Fuller, G.W., 2016. *New food product development: from concept to marketplace*, CRC Press.
- Ganglbauer, E. et al., 2009. Applying psychophysiological methods for measuring user experience: possibilities, challenges and feasibility. In *Workshop on user experience evaluation methods in product development*.
- Garrett, J.J., 2010. *Elements of User Experience, The: User-Centered Design for the Web and Beyond*, Pearson Education.
- Georges, V. et al., 2016. UX Heatmaps: Mapping User Experience on Visual Interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. pp. 4850–4860.
- Getrich, C.M. et al., 2016. Viewing Focus Groups Through a Critical Incident Lens. *Qualitative health research*, 26(6), pp.750–62.
- Gosling, S.D., Rentfrow, P.J. & Swann Jr, W.B., 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality*, 37(6), pp.504–528.
- Greig, A.D., Taylor, J. & MacKay, T., 2012. *Doing research with children: A practical guide*, Sage.
- Gubrium, J.F. & Holstein, J., 2012. SAGE: The SAGE Handbook of Interview Research: The Complexity of the Craft: Second Edition: : 9781412981644. In *The SAGE Handbook of Interview Research: The Complexity of the Craft*. pp. 27–44.
- Haddad, S., King, S, Osmond, P. and Heidari, S., 2012, November. Questionnaire design to determine children’s thermal sensation, preference and acceptability in the classroom. In *PLEA2012–28th Conference, Opportunities, Limits & Needs Towards an environmentally responsible architecture, 7–9 November 2012*.
- Hall, L. et al., 2005. Achieving empathic engagement through affective interaction with synthetic characters. *Affective computing and intelligent interaction*, pp.731–738.

- Hall, L., Woods, S., Dautenhahn, K., Sobral, D., Paiva, A., Wolke, D. and Newall, L., 2004, August. Designing empathic agents: Adults versus kids. In *International Conference on Intelligent Tutoring Systems* (pp. 604-613). Springer Berlin Heidelberg.
- Hall, L., Lufti, S., et al., 2011. Games based learning for Exploring Cultural Conflict. *AISB*.
- Hall, L. et al., 2014. Learning about cultural conflict through engaging with synthetic characters and cultures. *International Journal of Artificial Intelligence in Education*, 30(4), pp.415-440.
- Hall, L. et al., 2015. Learning to overcome cultural conflict through engaging with intelligent agents in synthetic cultures. *International Journal of Artificial Intelligence in Education*, 25(2), pp.291–317.
- Hall, L., Jones, S.J., Aylett, R., Hall, M., Tazzyman, S., Paiva, A. and Humphries, L., 2013. Serious game evaluation as a meta-game. *Interactive Technology and Smart Education*, 10(2), pp.130-146.
- Hall, L. et al., 2004. Using storyboards to guide virtual world design. In *Proceedings of the 2004 conference on Interaction design and children: building a community*. pp. 125–126.
- Hall, L. & Hume, C., 2012. Extending the Story into the Feedback Loop: Transmedia Evaluation. In *1st Global Conference Immersive Worlds and Transmedia Narratives*. Salzburg, pp36-42
- Hall, L. & Hume, C., 2011. Why Numbers, Invites and Visits are not Enough: Evaluating the User Experience in Social Eco-Systems. In *SOTICS 2011, The First International Conference on Social Eco-Informatics*. pp. 8–13.
- Hall, L., Hume, C. & Tazzyman, S., 2015. Engaging Children in Interactive Application Evaluation. *Special Issue on Children and Information Technology, Enfance*.
- Hall, L., Hume, C. & Tazzyman, S., 2016. Five Degrees of Happiness: Effective Smiley Face Likert Scales for Evaluating with Children. In *Proceedings of the The 15th International Conference on Interaction Design and Children*. pp. 311–321.
- Hall, L., Jones, S. & Aylett, R., 2011. Fostering Empathic Behaviour In Children And Young People: Interaction With Intelligent Characters Embodying Culturally Specific Behaviour In Virtual World. *INTED2011*.
- Hall, L., Jones, S. & Paiva, A., 2009. FearNot!: Providing Children with Strategies to Cope with Bullying. In *8th International Conference on Interaction Design and Children*. New York, NY, USA, pp. 276–277.

- Hall, L., Woods, S. & Dautenhahn, K., 2004. FearNot! Designing in the classroom. *British HCI, Leeds*.
- Hall, M., Hall, L., Hodgson, J., Hume, C. and Humphries, L., 2012, April. Scaffolding the Story Creation Process. In *CSEU (1)* (pp. 229-234).
- Hanna, L. & Ridsen, K., 1997. Guidelines for usability testing with children. *Interactions, Methods &*, pp.9–14.
- Haq, M., 2015. A Comparative Analysis of Qualitative and Quantitative Research Methods and a Justification for Adopting Mixed Methods in Social Research.
- Harlacher, J., 2016. An Educator's Guide to Questionnaire Development. REL 2016-108. *Regional Educational Laboratory Central*.
- Hartmann, J., Sutcliffe, A. and Angeli, A.D., 2008. Towards a theory of user judgment of aesthetics and user interface quality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 15(4), p.15.
- Hassenzahl, M., 2004. The Interplay of Beauty, Goodness, and Usability in Interactive Products. *Human-Computer Interaction*, 19(4), pp.319–349.
- Hastie, H. et al., 2016. I Remember You! Interaction with Memory for an Empathic Virtual Robotic Tutor. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*. pp. 931–939.
- Hazelden, K., 2007. An Extended Abstract for PRESENCE 2007, the 10 th Annual International Workshop on Presence. *Presence*, 2(8), pp.17–18.
- Hertzum, M., Molich, R. & Jacobsen, N.E., 2014. What you get is what you see: revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, 33(2), pp.144–162.
- Hofmann, S.G., Carpenter, J.K. & Curtiss, J., 2016. Interpersonal Emotion Regulation Questionnaire (IERQ): Scale Development and Psychometric Characteristics. *Cognitive Therapy and Research*, 40(3), pp.341–356.
- Horton, M., 2013. *Improving Validity and Reliability in Children's Self Reports of Technology Use*. phdthesis. University of Central Lancashire.
- Horton, M., Read, J.C. & Sim, G., 2011. Making your mind up?: the reliability of children's survey responses. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*. pp. 437–438.
- Howland, K., 2011. Designing an interface for multimodal narrative creation. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. pp. 1077–1080.

- Irvine, A., Drew, P. & Sainsbury, R.D., 2012. Am I not answering your questions properly? : Clarification, adequacy and responsiveness in semi-structured telephone and face-to-face interviews. *Qualitative Research*, 13(1), pp.87–106.
- Jäckle, A. & Eckman, S., 2016. *Is that still the same? Has that changed? On the accuracy of measuring change with dependent interviewing*,
- Jensen, J.J. & Skov, M.B., 2005. A review of research methods in children's technology design. In *Proceeding of the 2005 conference on Interaction design and children - IDC '05*. pp. 80–87.
- Johnson, K.A. et al., 2012. Using participatory scenarios to stimulate social learning for collaborative sustainable development. *Ecology and Society*, 17(2), p.9.
- Kano, A., Horton, M. & Read, J.C., 2010. Thumbs-up scale and frequency of use scale for use in self reporting of children's computer experience. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. pp. 699–702.
- Kaplan, B. & Maxwell, J., 2005. Qualitative Research Methods for Evaluating Computer Information Systems. *Evaluating the Organizational Impact of Healthcare Information Systems*, pp.30–55.
- Kaplan, J., 2015. Questionnaires. *BetterEvaluation.org*. Available at: <http://betterevaluation.org/evaluation-options/questionnaire> [Accessed August 29, 2016].
- Khanum, M.A. & Trivedi, M.C., 2012. Take Care: A Study on Usability Evaluation Methods for Children. Available at: <http://arxiv.org/abs/1212.0647> [Accessed September 9, 2016].
- Kincaid, J.P. et al., 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*,
- Kitchenham, B. & Pfleeger, S.L., 2002. Principles of survey research part 4: questionnaire evaluation. *ACM SIGSOFT Software Engineering Notes*, 27(3), pp.20.
- Kitzinger, J., 1995. Qualitative research. Introducing focus groups. *BMJ: British medical journal*, 311(7000), p.299.
- Kitzinger, J. & Barbour, R., 1999. Introduction: the challenge and promise of focus groups. *Developing focus group research: Politics, theory and practice*, pp.1–20.
- Klimmt, C. et al., 2012. Forecasting the Experience of Future Entertainment Technology: “Interactive Storytelling” and Media Enjoyment. *Games and Culture*, 7, pp.187–208.

- Krosnick, J.A., 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3), pp.213–236.
- Krosnick, J.A. et al., 2015. The future of survey research: challenges and opportunities. *The National Science Foundation Advisory Committee for the Social, Behavioral and Economic Sciences Subcommittee on Advancing SBE Survey Research*.
- Krosnick, J.A., 2000. The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Newsletter*, 20, pp.4–8.
- Krosnick, J.A., Narayan, S. & Smith, W.R., 1996. Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996, pp.29–44.
- Krueger, R.A. & Casey, M.A., 2014. *Focus groups: A practical guide for applied research*, Sage publications.
- Kruger, R.M., Gelderblom, H. & Beukes, W., 2016. The value of comparative usability and UX evaluation for e-commerce organisations.
- Kusunoki, D. & Sarcevic, A., 2012. Applying participatory design theory to designing evaluation methods. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. pp. 1895–1900.
- Laerhoven, H.V., Zaag-Loonen, H.V.D. and Derkx, B.H., 2004. A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta paediatrica*, 93(6), pp.830-835.
- Larsen, K.R., Nevo, D. & Rich, E., 2008. Exploring the semantic validity of questionnaire scales. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.
- Lavie, T. & Tractinsky, N., 2004. Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies*, 60(3), pp.269–298.
- Law, E., 2011. The measurability and predictability of user experience. *Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems EICS 11*, 29, pp.1–9.
- Likert, R., 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140, pp.55.
- Lim, K. et al., 2011. Technology Enhanced Role Play for Social and Emotional Learning Context - Intercultural Empathy. *Special Issue Journal of Entertainment Computing*, 2(4), pp.223–231.
- Linbo, L. et al., 2015. A review of interactive narrative systems and

- technologies: a training perspective. In *Simulation*.
- Lindström, M. et al., 2006. Affective diary: designing for bodily expressiveness and self-reflection. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*. pp. 1037–1042.
- Read, J.C. and Markopoulos, P., 2014. *Evaluating children's interactive products* (pp. 1043-1044). ACM.
- Marshall, C., 2016. Face-to-Face Interviews - Advantages and Disadvantages | Charlie Marshall | LinkedIn. *LinkedIn pulse*. Available at: <https://www.linkedin.com/pulse/face-to-face-interviews-advantages-disadvantages-charlie-marshall> [Accessed September 8, 2016].
- Matlin, M., 1994. *Cognition* 4th ed., Harcourt Press.
- Matson, J.L., Rotatori, A.F. & Helsel, W.J., 1983. Development of a rating scale to measure social skills in children: The matson evaluation of social skills with youngsters (MESSY). *Behaviour Research and Therapy*, 21, pp.335–340.
- McCambridge, J., Witton, J. & Elbourne, D.R., 2014. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology*, 67(3), pp.267–277.
- Melián-González, S., 2016. An extended model of the interaction between work-related attitudes and job performance. *International Journal of Productivity and Performance Management*, 65(1), pp.42–57.
- Mellor, D. & Moore, K.A., 2014. The use of likert scales with children. *Journal of Pediatric Psychology*, 39, pp.369–379.
- Merriam, S.B. & Tisdell, E.J., 2015. *Qualitative research: A guide to design and implementation*, John Wiley & Sons.
- Mundell, C., Vielma, J.P. & Zaman, T., 2016. Predicting Performance Under Stressful Conditions Using Galvanic Skin Response. *arXiv preprint arXiv:1606.01836*.
- Murchú, N.O., 2016. A designerly way of curating: reflecting on interaction design methods for curatorial practice. In *Curating the Digital*. Springer, pp. 9–19.
- Nacke, L.E., 2015. Games User Research and Physiological Game Evaluation. In R. Bernhaupt, ed. *Game User Experience Evaluation*. Human–Computer Interaction Series. Cham: Springer International Publishing, pp. 63–84.
- Nielsen, J. & Molich, R., 1990. Heuristic Evaluation of User Interfaces. In *ACM CHI'90 Conference (Seattle, WA, 1-5 April)*,. pp. 249–256.

- Nielsen, J., 1994. Heuristic evaluation. *Usability inspection methods*, 17(1), pp.25-62.
- Nielsen, J., 1997. The use and misuse of focus groups. *Software, IEEE*, 14(1), pp.94–95.
- Niyonzima, E., 2015. Perceptions and Experiences of Former Unaccompanied Refugee Children and Social Workers at a Care Home in Sweden.
- O'Brien, H.L. and MacLean, K.E., 2009, April. Measuring the user engagement process. In *Digital Life New World Conference, Boston, MA*.
- O'Brien, H. & Toms, E., 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1), pp.50–69.
- O'Brien, H.L., 2010. The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with Computers*, 22(5), pp.344–352.
- Oerke, B. & Bogner, F.X., 2011. Social Desirability, Environmental Attitudes, and General Ecological Behaviour in Children. *International Journal of Science Education*, pp.1–18.
- Ólafsson, K., Livingstone, S. & Haddon, L., 2013. Children's use of online technologies in Europe: a review of the European evidence base.
- Opendakker, R., 2006. Advantages and Disadvantages of Four Interview Techniques in Qualitative Research. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 7(4).
- Oppenheim, A.N., 1992. *Questionnaire design, interviewing and attitude measurement*, Continuum.
- Pasin, M. et al., 2015. A methodological approach to user evaluation and assessment of a virtual environment hangout. *Intelligent Technologies for Interactive Entertainment (INTETAIN), 2015 7th International Conference on*, pp.120–124.
- Portell, M. et al., 2015. Guidelines for reporting evaluations based on observational methodology. *Psicothema*, 27(3), pp.283–289.
- Portnoy, D.B. et al., 2008. Computer-delivered interventions for health promotion and behavioral risk reduction: a meta-analysis of 75 randomized controlled trials, 1988–2007. *Preventive medicine*, 47(1), pp.3–16.
- Qiong, X., 2015. Examining user engagement attributes in visual information search. In *iConference 2015 Proceedings*.

- Quax, P. et al., 2013. An evaluation of the impact of game genre on user experience in cloud gaming. In *2013 IEEE International Games Innovation Conference (IGIC)*. IEEE, pp. 216–221.
- Raita, E., 2012. User Interviews Revisited: Identifying User Positions and System Interpretations. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*. NordiCHI '12. New York, NY, USA: ACM, pp. 675–682.
- Rajeshkumar, S., Omar, R. & Mahmud, M., 2013. Taxonomies of User Experience (UX) evaluation methods. In *International Conference on Research and Innovation in Information Systems, ICRIIS*. pp. 533–538.
- Rawassizadeh, R. et al., 2015. Lesson Learned from Collecting Quantified Self Information via Mobile and Wearable Devices. *Journal of Sensor and Actuator Networks*, 4(4), pp.315–335.
- Read, J., MacFarlane, S. & Casey, C., 2002a. Endurability, engagement and expectations: Measuring children's fun. In *Interaction Design and Children*. Shaker Publishing Eindhoven, pp.189–198.
- Read, J., MacFarlane, S. & Casey, C., 2002b. Endurability, engagement and expectations: Measuring children's fun. *Interaction Design and Children*, 2, pp.1–23.
- Read, J.C. & MacFarlane, S., 2006. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. *Proceeding of the 2006 conference on Interaction design and children IDC 06*, pp.81.
- Reynolds-Keefer, L. et al., 2009. Validity issues in the use of pictorial Likert scales. *Studies in Learning, Evaluation Innovation and Development*, 6, pp.15–24.
- Reynolds-Keefer, L., Johnson, R. & Carolina, S., 2011. Is a picture worth a thousand words? Creating effective questionnaires with pictures. *Practical Assessment, Research & Evaluation*, 16, pp.1–7.
- Rice, M. et al., 2012. Co-creating games through intergenerational design workshops. In *Proceedings of the Designing Interactive Systems Conference*. pp. 368–377.
- Rizvic, S. et al., 2012. Interactive digital storytelling in the sarajevo survival tools virtual environment. In *Proceedings of the 28th Spring Conference on Computer Graphics - SCCG '12*. New York, New York, USA: ACM Press, pp. 109–116.
- Robertson, J., 2012. Likert-type scales, statistical methods, and effect sizes. *Communications of the ACM*, 55(5), p.6.
- Roto, V. et al., 2013. About «All About UX. *All About UX*. Available at: <http://www.allaboutux.org/about> [Accessed October 4, 2016].

- Sanoff, H., 2016. *Visual Research Methods in Design (Routledge Revivals)*, Routledge.
- Sas, C. et al., 2013. AffectCam: arousal-augmented sensecam for richer recall of episodic memories. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. pp. 1041–1046.
- Saunders, M., Lewis, P. & Thornhill, A., 2009. *Research Methods for Business Students*,
- Schoenau-Fog, H., 2011. Hooked!—Evaluating Engagement as Continuation Desire in Interactive Narratives. *Interactive Storytelling*, pp.219–230.
- Schoenau-Fog, H., 2012. Teaching Serious Issues through Player Engagement in an Interactive Experiential Learning Scenario. *Eludamos. Journal for Computer Game Culture*, 6(1), pp.53–70.
- Scriven, M., 2015. Evaluation Revolutions. *Journal of MultiDisciplinary Evaluation*, 11(25), pp.14–21.
- Scriven, M., 1991. *Evaluation thesaurus*, SAGE Publications, Incorporated.
- Segel, E. & Heer, J., 2010. Narrative visualization: Telling stories with data. *{IEEE} Transactions on Visualization and Computer Graphics*, 16(6), pp.1139–1148.
- Seitz, S., 2016. Pixilated partnerships, overcoming obstacles in qualitative interviews via Skype: a research note. *Qualitative Research*, 16(2), pp.229–235.
- Shapka, J.D. et al., 2016. Online versus in-person interviews with adolescents: An exploration of data equivalence. *Computers in Human Behavior*, 58, pp.361–367.
- Sharafi, P., Hedman, L. & Montgomery, H., 2006. Using information technology: engagement modes, flow experience, and personality orientations. *Computers in Human Behavior*, 22(5), pp.899–916.
- Short, C. et al., 2015. Designing engaging online behaviour change interventions: A proposed model of user engagement. *European Health Psychologist*, 17(1), pp.32–38.
- Silverman, D., 2016. *Qualitative research*, Sage.
- Simonsen, J. and Robertson, T. eds., 2012. *Routledge international handbook of participatory design*. Routledge.
- der Sluis, F., Van Dijk, E. & Perloy, L.M., 2015. Measuring fun and enjoyment of children in a museum: Evaluating the Smileyometer.
- Spool, J., 2015. Is Design Metrically Opposed? *UIE.com*. Available at:

https://www.uie.com/jared-live/transcripts/Is_Design_Metrically_Opposed.html [Accessed September 2, 2016].

- Stewart, D.W. & Shamdasani, P.N., 2014. *Focus groups: Theory and practice*, Sage Publications.
- Sutcliffe, A. & Hart, J., 2013. Some Reflections on Evaluating Users' Experience. In *which is commonly known as TwinTide (Towards the Integration of Trans-sectorial IT Design)*. pp.67–71.
- Symonds, J.E. & Gorard, S., 2008. The death of mixed methods: Research labels and their casualties. In *British Educational Research Association (Ed.), BERA Annual Conference, Heriot Watt University, Edinburgh*.
- Tasci, A.D. and Ko, Y.J., 2016. A fun-scale for understanding the hedonic value of a product: The destination context. *Journal of Travel & Tourism Marketing*, 33(2), pp.162-183.
- van Teijlingen, E. & Hundley, V., 1998. The importance of pilot studies. *Nursing standard: official newspaper of the Royal College of Nursing*, 16(40), pp.33–36.
- Timpany, G., 2011. Structured vs. Unstructured Questions. *Inquisium*. Available at: <http://survey.cvent.com/blog/customer-insights-2/structured-vs-unstructured-questions> [Accessed September 4, 2016].
- Tj, I., Leister, W., Schulz, T. and Larssen, A., 2015, May. The role of emotion and enjoyment for QoE—A case study of a science centre installation. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on* pp.1-6. IEEE.
- Traynor, M., 2015. Focus group research. *Nursing Standard*, 29(37), pp.44–48.
- Trochim, W.M.K. & Donnelly, J.P., 2006. *The Research Methods Knowledge Base*.
- Usability.gov, 2013. Usability.gov.
- Valstar, M.F., Almaev, T., Girard, J.M., McKeown, G., Mehu, M., Yin, L., Pantic, M. and Cohn, J.F., 2015, May. Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on* (Vol. 6, pp. 1-8). IEEE.
- Vannette, D. & Krosnick, J., 2014. A comparison of Survey Satisficing and Mindfulness. In *The Wiley Blackwell Handbook of Mindfulness*. pp.312.
- Vigo, M., Brown, J. & Conway, V., 2013. Benchmarking web accessibility

- evaluation tools: measuring the harm of sole reliance on automated tests. *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pp.1–10.
- Wade, A. & Demb, A., 2009. A conceptual model to explore faculty community engagement. *Michigan Journal of Community Service Learning*, 15(2).
- Wang, C.K.J. et al., 2008. Passion and intrinsic motivation in digital gaming. *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society*, 11(1), pp.39–45.
- Wang, G. et al., 2015. Constructive Play: Designing for Role Play Stories with Interactive Play Objects. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*. pp. 575–580.
- Weber, R. et al., 2009. Theorizing Flow and Media Enjoyment as Cognitive Synchronization of Attentional and Reward Networks. *Communication Theory*, 19(4), pp.397–422.
- Wengraf, T., 2003. Qualitative Research Interviewing: Biographic Narrative and Semi-Structured Methods. *Human Resource Development Quarterly*, 14(1), pp.117–122.
- Wharton, C. et al., 1994. The cognitive walkthrough method: A practitioner's guide. *Usability inspection*, pp.105–140.
- Wiemeyer, J. et al., 2016. Player Experience. In *Serious Games*. Springer, pp. 243–271.
- Wilson, C., 2014. *Interview techniques for UX practitioners: a user-centered design method*, Morgan Kaufmann.
- Wilson, J.R. & Sharples, S., 2015. *Evaluation of human work*, CRC Press.
- Wolff, W., Sandouqa, Y. & Brand, R., 2016. Using the simple sample count to estimate the frequency of prescription drug neuroenhancement in a sample of Jordan employees. *International Journal of Drug Policy*, 31, pp.51–55.
- Woodcock, B.A., 2014. "The Scientific Method" as Myth and Ideal. *Science & Education*, 23(10), pp.2069–2093.
- Woolrych, A. et al., 2003. Changing Analysts' Tunes: The Surprising Impact of a New Instrument for Usability Inspection Method Assessment. *People and Computers XVII: Designing for Society (Proceedings of HCI 2003)*.
- Wu, Z. et al., 2015. Effects of Display Characteristics on Presence and Emotional Responses of Game Players. *Human Behavior, Psychology, and Social Interaction in the Digital Era*, p.130.

- Wyse, S., 2014. Advantages and Disadvantages of Face-to-Face Data Collection. Available at: <http://www.snapsurveys.com/blog/advantages-disadvantages-facetoface-data-collection/>
- Yannakakis, G.N., 2008. How to model and augment player satisfaction: A review. In *Proceedings of the 1st Workshop on Child, Computer and Interaction, Chania, Crete, ACM Press*. pp. 1–5.
- Yannakakis, G.N. et al., 2013. Player Modeling. *Dagstuhl Follow-Ups*, 6, pp.59.
- Zaman, B. et al., 2012. Editorial: The evolving field of tangible interaction for children: The challenge of empirical validation. *Personal and Ubiquitous Computing*, 16, pp.367–378.
- Zaman, B., Vanden Abeele, V. & De Grooff, D., 2013. Measuring product liking in preschool children: An evaluation of the Smileyometer and This or That methods. *International Journal of Child-Computer Interaction*, 1, pp.61–70.
- Zikmund, W. et al., 2012. *Business research methods*, Cengage Learning.
- Zumbrunn, S., Marrs, S. & Mewborn, C., 2016. Toward a better understanding of student perceptions of writing feedback: a mixed methods study. *Reading and Writing*, 29(2), pp.349–370.

Appendix Content

- A. Context of Research
- B. Child & Parent consent form
- C. Parent/Guardian information Sheet
- D. Child information sheet
- E. Original version of the MESSY Scale (REF)
- F. Original version of the CQS
- G. Original version of Bryant's Empathy Index
- H. Review of children's media
- I. Workbook One
- J. Workbook Two
- K. Workbook Three
- L. Papers relating to this research

APPENDIX A: CONTEXT OF RESEARCH

In this section the practical backdrop of this research is provided. The aim of this chapter is to both contextualise this research and to clarify the distinction between eCute and the work that forms this PhD by providing the following information:

A.1 This research and the eCute Project: This section clarifies the association with and the distinction between the work that forms this research and the work that formed the outcomes of the eCute project, demonstrating that there is no overlap between the two and that the only contribution of eCute to this research was the provision of an evaluand and a set of evaluation requirements.

A.2 The eCute Project: This section introduces the eCute project, the European Union technology enhanced learning project that provided the context for this exploration of the evaluation of interactive narrative based learning applications, its project partners and research aims.

A.3 MIXER: Describes the MIXER application, characters (red and yellow teams) and narrative. An explanation of the rule sets used in MIXER is included. A description of the MIXER interaction modality concludes the section.

A.4 MIXER Evaluation: Outlines the eCute R&D team requirements for the MIXER Evaluation. Detailing the emotional, cognitive, behavioural and experiential goals of the MIXER application. The near and far transfer of learning is discussed and the validated questionnaires selected to assess the learning and engagement goals are detailed.

A1: This Research and the eCute Project

In order to carry out this research it was necessary to have something to evaluate (an evaluand), around which to develop and test the hypothesis and evaluation materials that form this research. My involvement with the eCute project, (as a research assistant), was serendipitous, as this involvement provided easy access to an application to evaluate, this was the MIXER application. Had I not been involved in the project the alternative would have been to find and approach a suitable research group in the hope that they would support me in my research by allowing access to whatever application they were developing. Fortunately this was not necessary.

The eCute project also had a set of evaluation requirements that I chose to incorporate into this PhD. The alternative would have been to create/invent a set of evaluation requirements, while this may have been an easier option, it was obvious that the inclusion of a set of complex, large scale and authentic evaluation requirements and constraints from a real research project offered the opportunity to add greater credibility and validity to this PhD.

As shown in the diagram below, the only contribution made to this research by the eCute Project was the provision of the MIXER application for use as an evaluand and the incorporation of the MIXER evaluation requirements, see appendix A3 for a description of MIXER and A4 for a description of the MIXER evaluation requirements.

The research aims of this PhD were outside and beyond the scope of the evaluation requirements of the eCute project. Once the final evaluation was complete the user data required by eCute (as detailed in Appendix A4) was handed over for analysis by eCute and the meta-evaluation data (e.g. completion rates, variance etc. see chapter 6 for a detailed discussion) was analysed for this research.

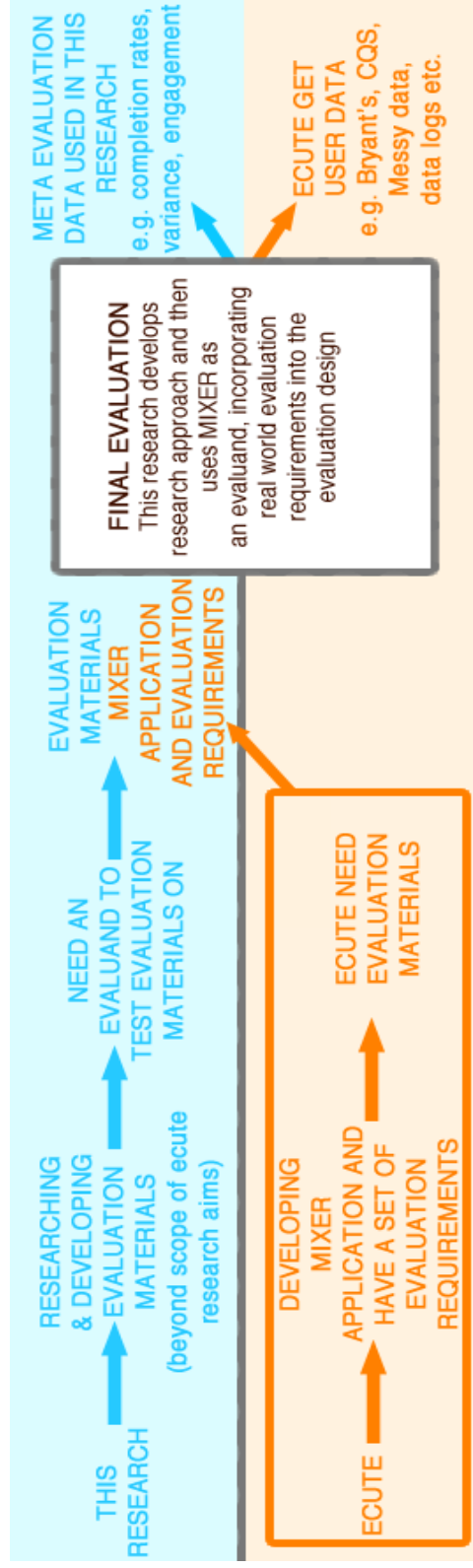


Figure A.1: Distinction between this research and eCute

A2: The eCute Project

eCute (education in Cultural understanding, technology enhanced) (www.ecute.eu) was funded by the European Union's Seventh Framework Programme for research and technological development (ICT-5-4.2 257666) and ran for 42 months from 2010-13. eCute was a technology enhanced learning project and the project focused on providing experiential learning for intercultural sensitivity using artificial intelligence techniques. The learning experience is provided through interacting with intelligent graphically embodied agents in a 3D virtual story world (see figure A.2).

eCute involved eight international partners, with the R&D team including over 25 researchers based at: Herriot Watt University, INESC-ID, University of Sunderland, Augsburg University, Wageningen University, Jacobs University, Seikei University and Kyoto University. The R&D team included computer scientists, interaction technologists, psychologists, information scientists, educational staff and evaluators, each with their own interests, requirements and constraints for evaluation.

eCute applied agent and interaction technologies to enhance intercultural sensitivity learning. Intelligent agents were developed through extending the Fatima architecture (Dias et al. 2011) creating a cultural agent architecture based on Hofstede's Cultural Dimensions (Hofstede et al. 2010) and Bennett's developmental model of intercultural sensitivity (Bennett 1986). To illustrate the potential of eCute's technology and approach, two showcases were developed: MIXER (Moderating Interaction for Cross Cultural Empathic Relations) (Aylett et al., 2014; Nazir et al., 2012) (see section 3.3) and TRAVELLER (Degens et al., 2013; Hall, Aylett, Hume, Krumhuber, & Degens, 2012), providing educational, innovative, narrative based, interactive applications to help develop cultural understanding and sensitivity in children and young adults.



Figure A.2: Screenshots taken from MIXER and TRAVELLER

eCute required a significant, large-scale evaluation of MIXER with multiple educational, technical and user experience evaluation requirements for an interdisciplinary project team. This thesis focuses on the evaluation of MIXER, with MIXER used as the evaluand around which the evaluation studies of this PhD were designed and delivered. The selection of MIXER as the evaluand in this PhD was related to two key factors: the timeliness and availability of the evaluand coupled with the importance, scale and complexity of the evaluation.

A3: THE MIXER APPLICATION

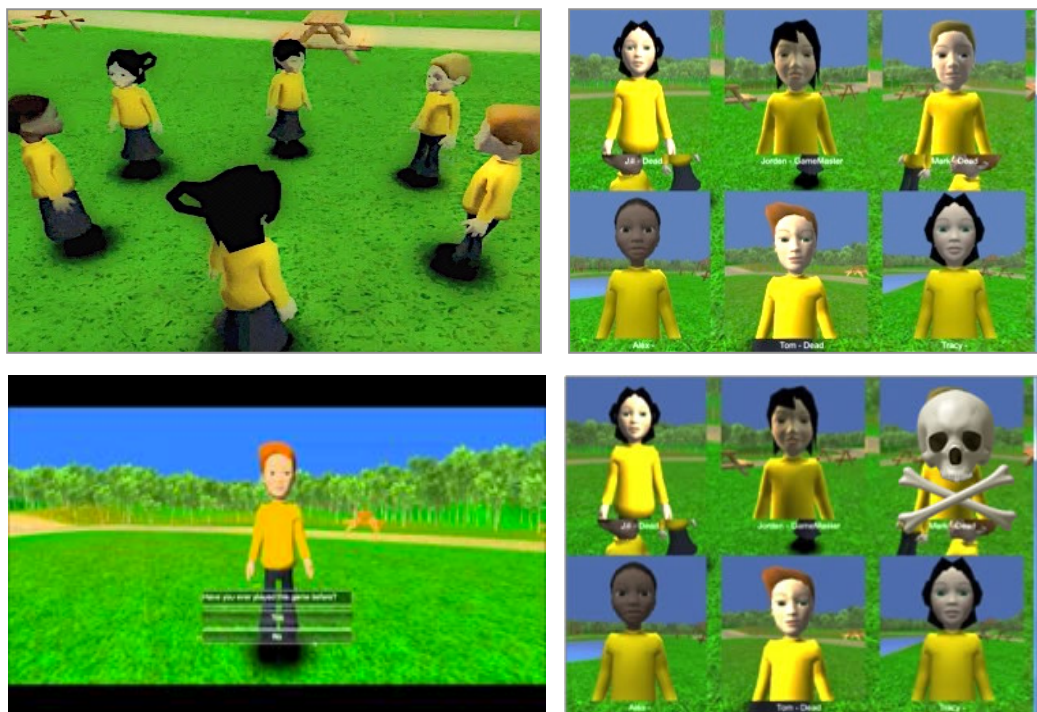


Figure A.3: Screen shots from the MIXER application

MIXER, the evaluand used in this research, is a computer-based intercultural sensitivity learning experience targeted at 9 - 11 year olds for use in the classroom context. MIXER was the first showcase to be developed in the eCute project and required a large-scale summative evaluation of a range of R&D issues with a minimum of 200 participants. MIXER aimed to provide an engaging user experience and the evaluation included assessing MIXER's success in engaging users in learning and interaction.

MIXER Narrative

MIXER engages users in an interactive narrative set in a virtual summer camp where two groups of school children (intelligent virtual agents) play Werewolves (Pallieres & Marly 2015), a popular intergenerational game widely known in many cultures. Werewolves' is a strategy-based turn-taking game, where participants adopt the roles of villagers, werewolves and narrator. The werewolf aims to 'kill' all of the villagers while the villagers try to identify the werewolf to end the game (see Aylett et al., 2014, for a detailed description of the simplified version of the werewolves game as implemented in the MIXER application).

MIXER aims to enhance learning about cultural conflict, and as such it depicts a peer conflict scenario, occurring when Tom (protagonist) plays the Werewolves game with two different groups of children at a summer camp, the Yellows and the Reds. Each team is composed of six intelligent agents - a game-master, a werewolf and four villagers. Following the application of Hofstede's work in the eCute approach, these teams provide two cultures or moral circles, each with different values, represented as two different rule sets for playing Werewolves.

Team	Rule difference
Yellow Team Rules	Each player takes turns to say whom he or she think the werewolf is and why, the player with the most votes is then killed off and is out of the game.
Red Team Rules	One player states who they believe the werewolf to be, if they do not have majority agreement from the other players then they themselves are killed by the villagers and they are out of the game.

Table A.1: Rule difference between the yellow and red teams

Tom plays one game with the yellow team and then moves on to play with the red team, who play the team with the different rule set. The rule change leads to a conflict situation and Tom accuses the red team of cheating. Acting as an invisible friend the child user helps Tom to understand the rule change and resolve the conflict. The final scene of MIXER shows that Tom understands and accepts the rules of the Red team and Tom tells the child user that the red team rules sound "Pretty cool" and that he can't wait to try out the new rules. This reinforces the message that sometimes what may appear to be unfair or strange behaviour may actually be due to simple differences that can be easily resolved and can result in positive outcomes. This observation and interaction

provides the basis for accepting people belonging to a given out-group into one's own 'moral circle' (Hofstede et al. 2009).

MIXER User Experience

The development of MIXER was user-centred, with the aim being to provide an engaging user experience, where the child user empathised and cared about what happened in the interactive narrative. To reinforce the children's engagement (Hall, Tazzyman, et al., 2014) in MIXER, children interact with a character, Tom, operating as his 'invisible friend' advising him about how to play the Werewolves game and interact with the Reds and the Yellows. With the application of the similarity principle (Hall & Woods, 2006) the characters (see 3.5) were designed to be similar to the intended users, with age appropriate appearance and voices, including boys and girls with a mix of ethnicities. The two teams were dressed (as is common in summer camps), in team T-shirts, representing the Reds and the Yellows.

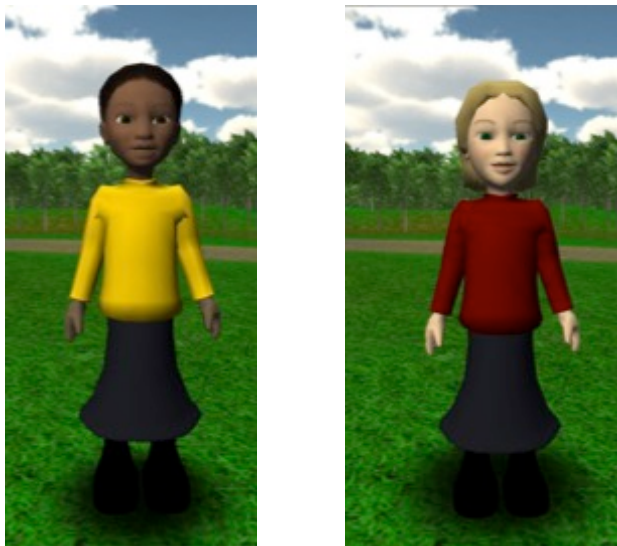


Figure A.4: Characters from the MIXER application

To further increase engagement, in MIXER, an innovative and exciting interaction approach was developed, with the child user interacting with Tom through a tablet, connected to a PC via Wi-Fi (see fig. A.5), using a Pictorial Interaction Language (Endrass et al. 2014), (see fig A.6).



Figure A.5: MIXER set up using iPad, PC and Wi-Fi connection

The Pictorial Interaction Language (see figure A.5) provides children with access to over 70 graphics structured for use in sentences, enabling them to create their advice for Tom. In addition to being fun, intuitive and engaging, this approach also reinforces the child's role as invisible friend. Whilst everyone can see the virtual environment (on the computer screen), only the child themselves can see their 'private' communication with Tom on the tablet.

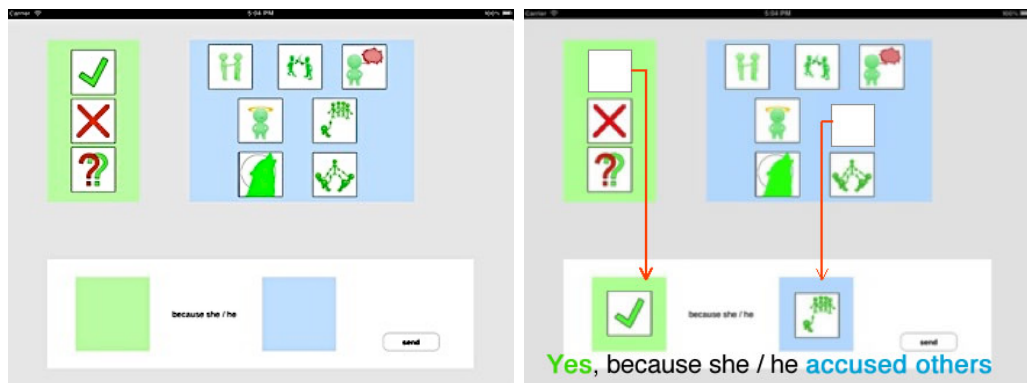


Figure A.6: Example screens from the Pictorial Interaction Language

A4: MIXER EVALUATION REQUIREMENTS

The evaluation requirements for the MIXER application were defined by the eCute project (The eCute Project 2011). In the large-scale summative evaluation of MIXER, the main aim was to identify if the following learning outcomes were achieved:

- **Emotional:** MIXER supports children to recognise emotions (for example fear and anxiety) when dealing with the strange behaviours of another group
- **Cognitive:** MIXER supports children to start learning the specific practices and values of another group
- **Behavioural:** MIXER supports children in being fully present in attending to others' verbal and non-verbal messages

In addition the following goal related to whether the MIXER technology was an effective approach for technology-enhanced learning:

- **Experience:** MIXER engages children in the narrative and with the characters, supporting the children's understanding and learning of strategies for coping with intercultural conflict

To establish if these learning goals were met, MIXER was to be evaluated in two ways: 1) Directly after the MIXER interaction aiming to identify if MIXER resulted in near transfer or immediate learning; and 2) With a pre- post- test design focusing on whether interacting with MIXER resulted in far transfer or sustained learning.

The first aim of the MIXER application was for children to show improvement against a set of cultural learning goals (see table A.2) for both near and far transfer, following eCute's cultural learning framework (Swiderska et al. 2011).

Attitude >	Emotional goals	Cognitive goals	Behavioural Goals
Stage of Learner v			
Beginner (conscious incompetence) Observation and acquisition	Be able to recognise your emotions when dealing with behaviour of another group	Start learning the specific practices and values of another group	Be fully present in attending to the other's verbal and non-verbal messages
Journeyman (conscious competence) Relating and experimenting	Be able to observe the behaviour of another group without feeling prejudice	Understand on a basic level the differences and similarities between another group and your own	Practise skills learned in the previous stage and experiment with different forms of behaviour
Expert (unconscious competence) Adapting and belonging	Be able to share emotions of a member of another group and other's experiences through empathy	Players should be able to discriminate and select appropriate strategies in cultural contexts.	Be able to unconsciously participate in a group as a native.

Table A.2: eCute's cultural learning framework

For the evaluation of far transfer, three validated instruments were selected by the psychology and educational partners in eCute to assess the learning goals as a pre- and post- measure. The instruments were Cultural Intelligence Scale (CQS) (Ang et al., 2007), Bryant's Empathy Index (Bryant 1982) MESSY (Matson Evaluation of Social Skills) (Matson et al., 2010) (see appendix E, F and G). The following table (A.3) shows the learning goal against the instrument selected and the rationale of use for each instrument.

Learning Goal	Instrument	Rationale
EMOTIONAL Be able to recognise emotions (e.g. fear and anxiety) when dealing with the novel / unknown behaviours of another group?	Cultural Intelligence Scale (CQS) (Ang et al., 2007)	The behavioural subscale of the CQS is used as a pre and post measure of a child's capability to adapt verbal and nonverbal behaviour in different situation/cultures. This will provide data for the question: "Do children who have a more flexible repertoire of behavioural responses in culturally diverse settings recognise more emotion/behaviours in the MIXER application?" This will address aspects of the behavioural and emotional learning outcomes.
COGNITIVE Start learning the specific practices and values of that group?	Bryant's Empathy Index (Bryant, 1982)	Factor One from the Bryant Empathy Index will be used as a measure of children's empathic behaviour and styles. This will provide data for the question: "Are children with higher empathy levels more able to recognise and accept emotions in novel situations?" This will address the emotional goal of the learning outcomes: "Be able to recognise your emotions when dealing with strange behaviours of another group".
BEHAVIOURAL Being fully present in attending to others verbal and non-verbal messages.	MESSY (Matson Evaluation of Social Skills) (Matson et al., 2010)	Factor two and four of the MESSY questionnaire have been selected to determine children's capability to adapt to verbal and nonverbal behaviour in different situations/cultures to assess the behavioural goal from the learning outcomes: "Be fully present in attending to others verbal and nonverbal messages".

Table A.3: MIXER learning goals, instruments and rationale

The evaluation of near transfer was assessed through the Experience Evaluation Questionnaire (EEQ). This instrument was developed to evaluate the user learning experience in VLEs populated by embodied characters, based on Hall et al., (2013) and Hall, Woods and Aylett, (2006). The EEQ collects data related to children's immediate learning, their narrative comprehension, and empathic engagement. The EEQ addresses all four MIXER learning goals: emotional, cognitive, behavioural and experiential, as detailed in table A.3.

In addition the EEQ met the second major aim of the MIXER evaluation investigating whether children found interaction with MIXER an engaging, interesting and enjoyable experience. Assessing engagement was also an important method of verifying any outcomes of the learning goals, i.e. if a child scored low on the learning goals it was hypothesised that this was because they were not engaged with the application for some reason. The following table details the engagement goals for MIXER.

	Engagement goals
Agents	All aspects of agent believability and effectiveness, both in terms of presentation, communication and mind architecture.
Engagement	The level of engagement experienced by the user in respect to their interaction with MIXER.
Comprehension	The level of the user's understanding of the events and progression of the scenario and interaction.

Table A.4: EEQ Evaluation Goals

Part of the purpose of the EEQ is to assess how engaged participants were with the characters, story, interaction approach and experience as detailed in the following table.

Issue	Evaluation
Character Preferences	Having used MIXER, children should have engaged with and have a deeper relationship with Tom than any of the other characters. This relates to the emotional and behavioural learning objectives, with the need to empathise with Tom and taking the role of invisible friend, key to learning.
Narrative Comprehension	Children's narrative comprehension of the MIXER scenario. Evaluation aims to assess whether children listened and paid attention to the story line and their degree of emotional, behavioural and cognitive learning. The children's opinions of the rule conflict were also evaluated to assess the cognitive and behavioural learning outcomes.
User Experience / Usability	Questions on user experience with MIXER (e.g. appropriateness of duration, desire to use MIXER again, etc.). Assessment of usability (e.g. voices, text, etc.) and experience (e.g. who explained the rules the best) of the MIXER application.
Interaction approach	Evaluation of the Pictorial Interaction Language – usability, user experience and enjoyment.

Table A.5: Experience Evaluation Questionnaire

Summary

This chapter has described how this research is separate from the goals and activities of the eCute project and contextualised the work presented in this thesis by explaining its position amongst the various elements (eCute, MIXER, etc.) that were required to carry out this research.

APPENDIX B: PARENT/GUARDIAN AND CHILD CONSENT FORM



Parent/Guardian and Child Consent Form

Study Title: INSERT STUDY NAME - **eCUTE Project**

Name [Parent or guardian]

Child's Name:

Address:
.....

I have discussed this study with my child/children and I give consent for my child/children to be a participant. We have both been informed about what participation will involve and I understand that I can withdraw my child/children at any time without giving reason and without penalty.

I give consent for my child/children's supplied data to be discussed by research workers in the study, online and to be used for research dissemination.

I also give consent for photographs and videos to be made relating to my child's/children's participation and understand that these will ONLY be used for research discussion and dissemination.

Photographs Yes

Video Yes

Signed
[Parent or Guardian]

Date:

This study has been approved by the University of Sunderland Ethics Committee



APPENDIX C: PARENT/GUARDIAN INFORMATION SHEET

eCUTE Project Participation

This study is part of a European project called eCUTE, which focuses on enhancing learning through the use of technology. As part of this project we are trying to develop new ways to evaluate interactive games and applications. However, our aim is not just to evaluate interactive applications, but also to understand how children and teenagers want to be evaluated. Evaluation is when we ask people what they think of a story, game or a piece of software.

The idea that we are studying with your child (ren)'s help is the use of an evaluation approach based on a child/participant-centered approach. During their participation in the project, they will use an iPad to control a game. At certain points we will ask your child(ren) to tell us about their experiences, ideas and thoughts about the sessions both in words and in pictures.

The results and outputs from this study will be discussed within the eCUTE Project and may be used for research dissemination. Taking part in the study is entirely voluntary and will be conducted over several sessions, your child (ren) are free to stop at any time. This study has been approved by the University of Sunderland Research Ethics Committee.

Contact Details for Further Information

If you have any questions about the study or issues you want to discuss, contact Lynne Hall [0191 515 3863].

You can also contact the Chairperson of the Research Ethics Committee of the University of Sunderland:

Dr. R Pullen

Chairperson of Research Ethics Committee

Faculty of Applied Sciences

University of Sunderland

Sunderland

Tel: 0191 515 2609

email: robert.pullen@sunderland.ac.uk

APPENDIX D: CHILD INFORMATION SHEET

The idea that we are studying, with your help, is the way you can help use, make and test new games for schools.

During your involvement with the project, you will be asked to use an iPad to control a PC and testing new software.

At different times we will ask you what you think of the things you have done.

Thank you,

The Research Team, University of Sunderland.



APPENDIX E: ORIGINAL FORMAT - MESSY SCALE

(Matson Evaluation of Social Skills with Youngsters)

Social Skills/Assertiveness Subscale

MESSY QUESTIONNAIRE: SOCIAL SKILLS/ASSERTIVENESS SUBSCALE

*The Matson Evaluation of Social Skills with Youngsters: Social
Skills/Assertiveness Subscale*

I help a friend who is hurt:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I cheer up a friend who is hurt:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I feel good if I help someone:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I ask if I can be of help:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I ask others how they are, what they have been doing etc.:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I do nice things for people who are nice to me:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I stick up for my friends:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I look at people when they are speaking:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I say 'thank you' and I am happy when someone does something for me:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I laugh at other people's jokes and funny stories:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I share what I have with others:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I know how to make friends:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I feel happy when someone else does well:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I ask questions when talking with others:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I feel sorry when I hurt someone:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I walk up to people and start a conversation:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I see my friends often:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I call people by their names:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I take care of other's property as if it were my own:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I show my feelings:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I keep secrets well:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I join in games with other children:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I look at people when I talk with them:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I explain things more than I need to:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

I make other people laugh:

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Strongly	Agree	Neutral	Disagree	Strongly
Agree				Disagree

APPENDIX F: ORIGINAL FORMAT - CQS - Cultural Intelligence Scale

The full version of the CQS is shown below, the Behavioral CQ sections used in the MIXER evaluation are shown in red:

Circle the answer that BEST describes you **AS YOU REALLY ARE**.

1 = strongly disagree; 7 = strongly agree

I am aware of the things I know about others I use when playing or working with people with different backgrounds.

1 2 3 4 5 6 7

I change what I believe about people as I work or play with people from a background that I don't know.

1 2 3 4 5 6 7

I am aware of what I know about people of different backgrounds and use this when I work or play with people of different backgrounds.

1 2 3 4 5 6 7

I check if I am right as I work and play with people from different backgrounds.

1 2 3 4 5 6 7

I know about the rules and money of people with different backgrounds.

1 2 3 4 5 6 7

I know the rules (e.g., vocabulary, grammar) of other languages.

1 2 3 4 5 6 7

I know the values and religious beliefs of people with a different background.

1 2 3 4 5 6 7

I know how people with a different background get married.

1 2 3 4 5 6 7

I know about arts and crafts of people with a different background.

1 2 3 4 5 6 7

I know why people's body movements are different when they are talking.

1 2 3 4 5 6 7

I enjoy playing or working with people from a different background.

1 2 3 4 5 6 7

I am confident that I can work or play with locals in a place that I don't know.

1 2 3 4 5 6 7

I am sure I can deal with the stresses of adjusting to a place that is new to me.

1 2 3 4 5 6 7

I enjoy living in places that are unfamiliar to me.

1 2 3 4 5 6 7

I am confident that I can adjust to the way people shop in a different place.

1 2 3 4 5 6 7

I change the way I talk (e.g., accent, tone) when working or playing with people from a different background.

1 2 3 4 5 6 7

I use pause and silence differently to suit different situations involving people from a different background.

1 2 3 4 5 6 7

I vary the rate of my speaking when dealing with people from a different background if a situation requires it.

1 2 3 4 5 6 7

I change the way I move my body when dealing with people from a different background if a situation requires it.

1 2 3 4 5 6 7

I alter my facial expressions when dealing with people from a different background if a situation requires it.

1 2 3 4 5 6 7

APPENDIX G: ORIGINAL FORMAT - Bryant's Empathy Index

Answer **Yes** or **No** to each of the questions.

- 1 It makes me sad to see a girl who can't find anyone to play with
 - 4 I really like to watch people open presents, even when I don't get a present myself
 - 5 Seeing a boy who is crying makes me feel like crying
 - 6 I get upset when I see a girl being hurt
 - 7 Even when I don't know why someone is laughing, I laugh too
 - 8 Sometimes I cry when I watch TV
 - 11 I get upset when I see an animal being hurt
 - 12 It makes me sad to see a boy who can't find anyone to play with
 - 13 Some songs make me so sad I feel like crying
 - 14 I get upset when I see a boy being hurt
 - 15 Grown-ups sometimes cry even when they have nothing to be sad about
 - 19 Seeing a girl who is crying makes me feel like crying
 - 22 I don't feel upset when I see a classmate being punished by a teacher for not obeying school rules
-

APPENDIX H: REVIEW OF CHILDREN'S MEDIA

In addition to consulting academic literature to gain an understanding of the theory behind evaluation, response bias and engagement, a review of media targeted to the age group of 9-11 year olds was also conducted. The aim of this review was to inform the design of the workbooks by understanding the design practice applied in the production of children's media. The reviewed media included both educational and recreational literature, as these are the two most frequently accessed forms of literature by 9 to 11 year olds. Educational media included various SATs and Key stage 2 support materials in both print and online. However, as fun and enjoyment were crucial design features in terms of engagement, recreational media was the main focus of the review, there was also a lot more variety in the recreational literature which included comic books / magazines, special interest magazines (e. g. Doctor Who, Bird Life etc.), activity books / sheets, websites, sticker collecting books, fiction and non-fiction books and annuals. The review focused on the following; firstly Content and Activities, it was important to understand what content and activities were fun for children. Fun and enjoyment were selected as a measure of the children's engagement with the evaluation materials. To design evaluation as a fun experience, (as similar as possible to the fun experienced when completing a recreational activity book), it was not only helpful and inspiring to look at age appropriate media but it was also essential that the designs of the workbooks were not based on assumptions of what children's media 'should contain' or how they 'should look'. The literature review highlighted several key areas that were to be further established via the review of children's media. These included the the need for variety in the levels of engagement experienced. Engagement should vary, for example, by starting low and building, then dropping back and then building again. This pattern of peaks and troughs should be evident in the placement of activities in the workbooks and aligns with (Hanna & Ridsen, 1997) recommendation that children's focus when completing demanding activities should be limited to short five to ten minute intervals, combined with breaks in activity intensity. It was intended to break up evaluation activities in the workbooks with filler activities and it was anticipated that the review of children's media would verify these design decisions. Secondly, Layout and Aesthetics, elements such as colours, fonts and use of images etc. were

reviewed for similar reasons. It would be easy to assume that primary colours and juvenile fonts (e. g. Bradley or comic sans) would be a suitable choice in the design of childrens media, it was important and worthwhile to give a deeper consideration to these elements of the design. Finally, the use of characters and narrative applied throughout the literature (i. e. animals or characters from TV/film) was also reviewed. Narrative and character engagement form a crucial part of the Transmedia Evaluation methodology (see section 3.2), and were also one of the eCute evaluation requirements for MIXER (see section 3.4).

Content and Activities

While the content in the media reviewed varied depending on the topic of the publication, there was a trend towards providing a mix of informational features and entertainment. In addition to the subject specific articles, i.e. Moths in Bird Life magazine [ref] and Justin Bieber in Top of the Pops [ref], there were also activities such as word searches, mazes, spot the difference and colouring activities [e. g. where examples can be found]. These activities were interspersed between articles, providing variety to the reader to maintain engagement. Although many of the publications reviewed were not traditional comic books, there were comic strips used as a story telling mechanism in the majority of the publications.



Figure H.1: Example comic strip from Bird Life magazine

Stickers were a found frequently in many of the publications. In the form of sticker collecting books (e. g. Panini) or free stickers with comics and magazines and some books had special plastic coated pages so that stickers could be added and repositioned.



Figure H.2: Examples of sticker book activities

Quizzes were another popular activity. While the subjects varied the format was frequently the same, children answer questions and then calculate if they were mostly A, B or C before reading a description of their result at the bottom of the page, (see figure 3). The quiz results/outcomes were generic descriptions intended, in general, to make the reader feel happy about themselves and their choices. The question and answer format is similar to questionnaires but the quizzes provide immediate feedback to the reader.



Figure H.3: Example of quiz feedback

Layout

The layouts used in the paper based publications versus an online/digital format had obvious differences, for example, digital applications are able to make use of features such as 'next' buttons to help guide a user through an online experience. Producers of paper-based literature maintain engagement (novelty) by making the design of each page layout very different in appearance from the next.



Figure H.4: Producers maintain engagement / novelty by designing pages that vary in appearance

However, to ensure that the layout of each individual page is readable the designs contain an overall consistency, for example, by using borders and different coloured backgrounds to highlight the information in each section. At first glance the page (see figure 5) looks haphazard, but each section is laid out with a title, a piece of text and an image and this is consistent throughout this page.



Figure H.5: Layouts look casual and haphazard while maintaining consistent design Principles to aid the reader

Numbering (see figure 6) and arrows also help guide the reader from the start to the end of the article.



Figure H.6: Use of numbering to guide readers through the article (Taken from REF Top of the Pops, Issue 228, 12.09.12)

In the 'Epic Nature Spotter' activity below (McDonalds, 2013) the children are asked to follow a trail and count the wildlife that they spot to see which character spotted the most. The use of a curved line in this activity provides several benefits; firstly, the activity has a clear signposting of the start and end of the activity. This is very useful in an evaluation context, particularly in a classroom setting where the children need to complete the task with as little assistance as possible. The children can see where to start, where they need to get to and can complete the task as they go. Secondly, placing questions on a curved, rather than straight line may reduce occurrences of the straight-lining response bias.

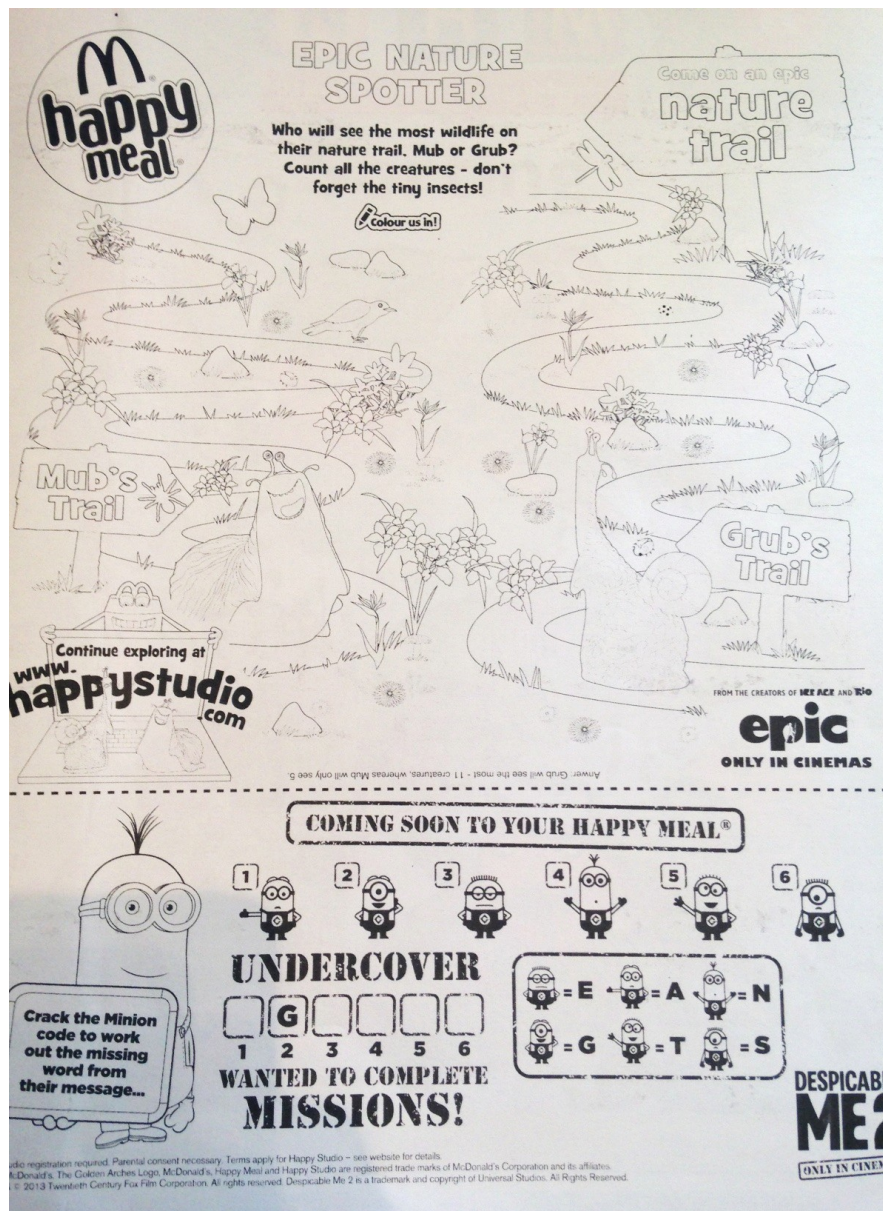


Figure H.7: McDonalds activity sheet

Aesthetic elements

As is typical with publications designed for a younger audience the use of colour is bold, bright, clashing and eye catching.



Figure H.8: Bright bold fonts and colours used

Similarly, wide ranges of fonts were applied within single page design layouts. Images and other decorative elements, such as boxed fonts (see figure H.8) borders and backgrounds were used on most pages. Where publications were targeting a specific gender, the use of pink and blue was dominant as is shown in figure H.9.



Figure H.9: Gender specific media for children and teenagers

Use of characters and narrative

The Doctor Who magazine was a great example of transmedia entertainment. Characters from the TV show were used throughout the magazine with features and activities about the various characters and plot lines. The magazine also carried the Doctor Who theme throughout the features and activities provided, maintaining a strong link between the onscreen experience and the paper based publication.



Figure H.10: Example of character usage

Lessons learnt for evaluation from children's media

Clearly, the producers of children's media understand how to engage children, with valuable lessons available for the designers of evaluations with children. After reviewing children's media it is evident that designing for engagement may also provide solutions to the existing issues relating to response bias by improving the user experience of evaluation. The following table provides a summary of the reviewed item, its possible purpose in reducing response bias in an evaluation context and where it was used in the design of the workbooks, along with the relevant section.

Media review item	Purpose	Used
Character / narrative theming throughout e.g. Doctor Who,	Introduces and reinforces character and scenario, embedding the evaluation into the narrative.	Tom is used throughout workbook 1 and 2. Workbook 3 was testing far transfer so Tom does not appear but is named once.
Filler activities – word searches, maze etc.	The media reviewed followed up a text heavy page with either a poster, a very visual picture or an activity	The workbooks follow a similar pattern of filler activity, followed by evaluation activity, another filler activity etc
Curved lines from McDonalds Epic Nature Spotter activity	Removing linearity from questionnaire format	Used throughout workbooks, New Friendzzz, which woodland animal, yes or no
Arrows and numbering to guide readers	Used to guide from start to end of activity so that no questions are missed – ensures 100% completion with minimal interruption/assistance	YES OR NO
Comic strips	Used in media to entertain, very visual way of story telling with small amounts of text reducing cognitive effort required	The Trip
Quizzes	Engaging way of asking questions	Used in 'Which woodland Animal are you?'
Stickers	Used as an alternative to pen/pencil. May delay response and encourage optimal response	Used in workbook 2 & 3, Purposefully not used in workbook one to offer novelty in workbook 2 Used in Who wins, True or false, and in Workbook 3 think fast page numbers

Table H.1: Summary of lessons learnt from reviewing children's media

APPENDIX I: WORKBOOK ONE

APPENDIX J: WORKBOOK TWO

APPENDIX K: WORKBOOK THREE

APPENDIX L: PUBLICATIONS RELATING TO THIS RESEARCH TO DATE

Hall, L., Hume, C. and Tazzyman, S., 2016, June. Five Degrees of Happiness: Effective Smiley Face Likert Scales for Evaluating with Children. In *Proceedings of the The 15th International Conference on Interaction Design and Children* (pp. 311-321). ACM.

Hall, L., Hume, C. and Tazzyman, S., 2015. Engaging Children in Interactive Application Evaluation. *Enfance*, 2015(01), pp.35-66.

Endrass, B., Hall, L., Hume, C., Tazzyman, S. and André, E., 2014, June. A Pictorial Interaction Language for Children to Communicate with Cultural Virtual Characters. In *International Conference on Human-Computer Interaction* (pp. 532-543). Springer International Publishing.

Hall, M., Hall, L., Hodgson, J., Hume, C. and Humphries, L., 2012, April. Scaffolding the Story Creation Process. In *CSEDU (1)* (pp. 229-234).

Hall, L. and Hume, C., 2011, October. Why numbers, invites and visits are not enough: Evaluating the user experience in social eco-systems. In *SOTICS 2011, the first international conference on social eco-informatics* (pp. 8-13).



M
MAG



QUIZZES
FUN ACTIVITIES
WORDSEARCH



The Super
Summer
Special

What's your name?

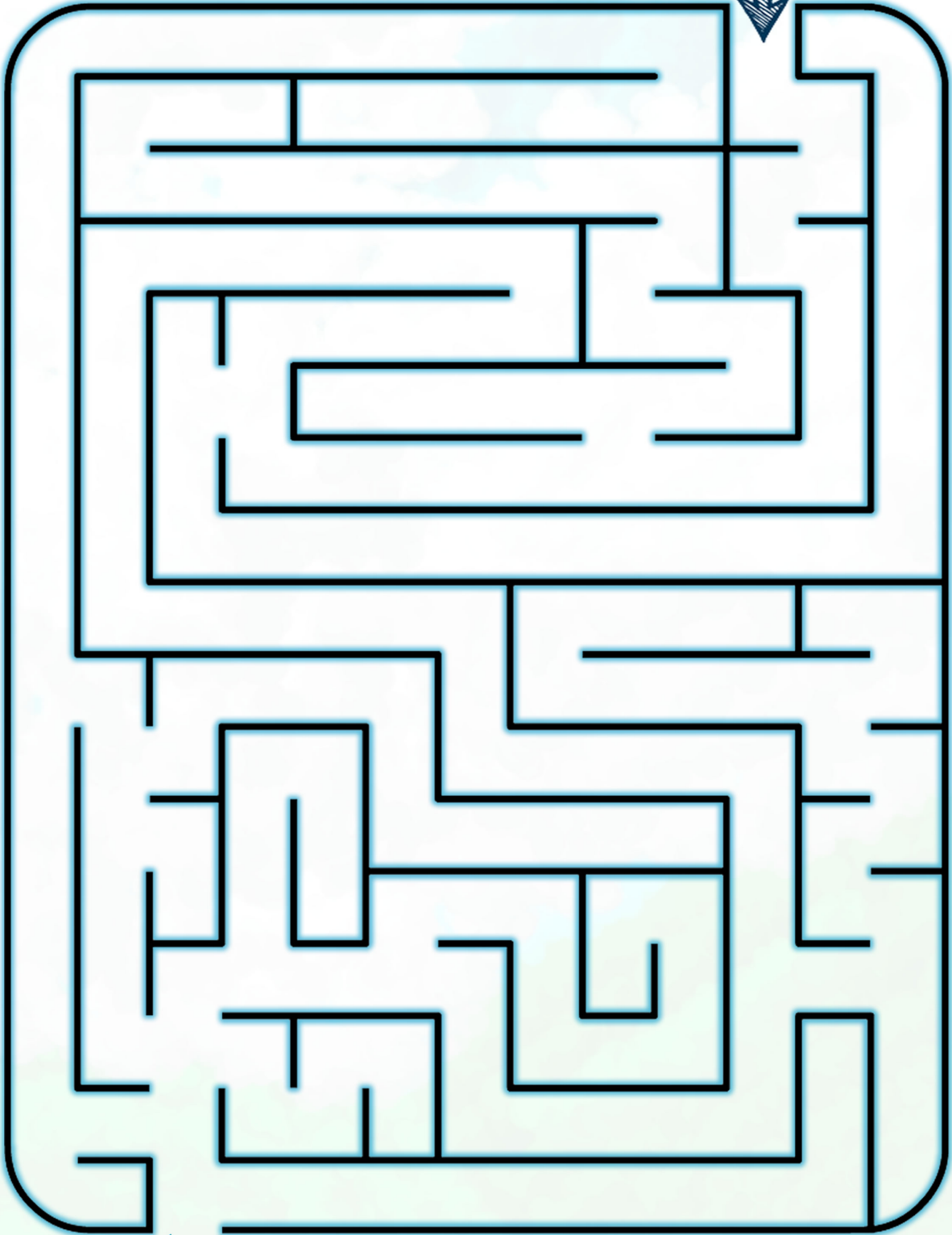
how old are you?

are you a boy or a girl?

FIND THE CAMP!!!



Can you help Tom to find his way back to camp before it rains?



New Friendzzz

It looks like Ben and Barney the bees could do with some help to become friends.

Help Ben by telling him about you.

Circle the face that best describes how much you match the statement.



A lot like you



Mostly like you



Like you



Sometimes like you



Not like you



I call people by their real name



I walk up to people and start a conversation



I show my feelings



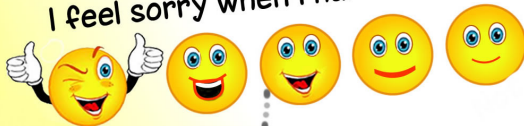
I know how to make friends



I look at people when they are speaking



I feel sorry when I hurt someone



I ask questions when talking with others



I see my friends often



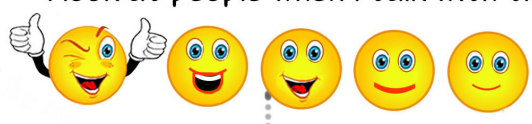
I cheer up a friend who is hurt



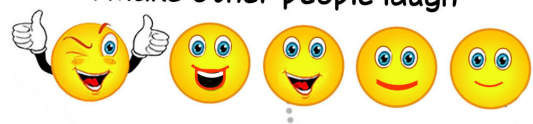
I stick up for my friends



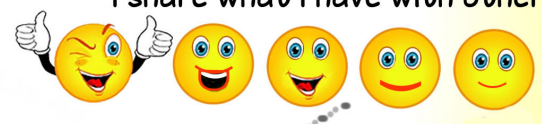
I look at people when I talk with them



I make other people laugh



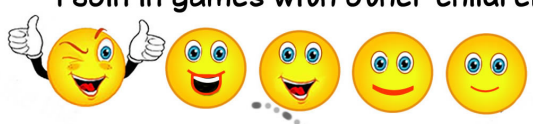
I share what I have with others



I laugh at other peoples jokes and funny stories

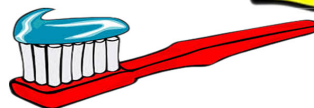
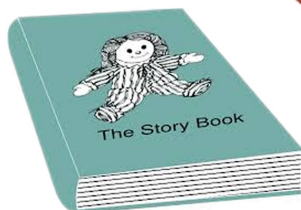
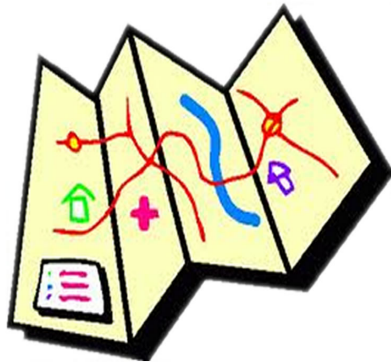


I join in games with other children



ALL PACKED & READY TO GO!

Draw a line from the object to the backpack to show the **five** things you would chose to take on a trip.



WHICH WOODLAND ANIMAL ARE YOU?

Take this quick quiz to find out if you're a Badger, Fox or stag.
Read each question and circle the smiley face that best shows how much it is or is not like you.



A lot like you



Mostly like you



Like you



Sometimes like you



Not like you

When you meet someone new do you change the way you talk?



When you meet someone new do you always slow down when you are speaking?



When you meet someone new do you speak more slowly with more pauses and spaces?



When you meet someone new do you change the way you move your body?



When you meet someone new do you change your facial expressions?



Your teacher will tell you which animal you are:

BADGER

Badgers like to be alone, but have lots of friends.

They can be fierce and will protect each other at all costs.

FOX

Foxes like to hangout together and have small groups of friends.

They are generally shy but will help others out if needed.

DEER

Deer like to have lots of friends but tend to fight with each other.

They tend to keep new deer away, but look after each other.

YES OR NO...

Tom is nervous.
It's the first time he's been
away from home.
Tom can see different things as he
waits for the bus.
Circle

YES or NO
to say if you agree with Tom.

IT'S HARD FOR ME TO
SEE WHY SOMEONE ELSE
GETS UPSET...

YES NO

PEOPLE WHO
KISS AND HUG
IN PUBLIC ARE SILLY!

YES NO

IT'S SILLY TO TREAT
CATS AND DOGS AS
THOUGH THEY HAVE FEELINGS
LIKE PEOPLE...

YES NO

KIDS WHO HAVE NO
FRIENDS PROBABLY DONT
WANT ANY...

YES NO

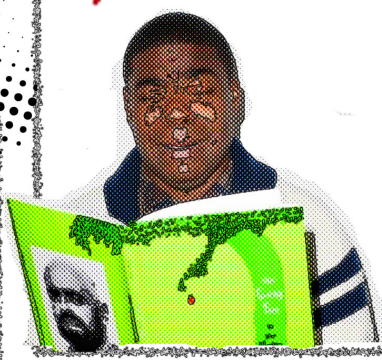
KIDS WHO CRY BECAUSE THEY ARE HAPPY ARE SILLY...



YES NO



I THINK IT'S FUNNY THAT SOME PEOPLE CRY DURING A SAD MOVIE OR WHILE READING A SAD BOOK



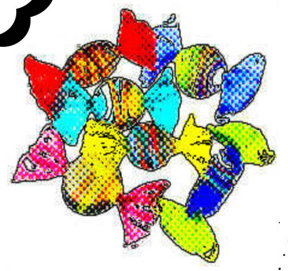
YES NO



I AM ABLE TO EAT ALL OF MY SWEETS EVEN WHEN I SEE SOMEONE LOOKING AT ME WANTING ONE...



YES NO



I GET MAD WHEN I SEE A CLASSMATE PRETENDING TO NEED HELP FROM THE TEACHER ALL THE TIME...



YES NO



I DONT FEEL UPSET IF A CLASSMATE IS PUNISHED FOR BREAKING THE RULES



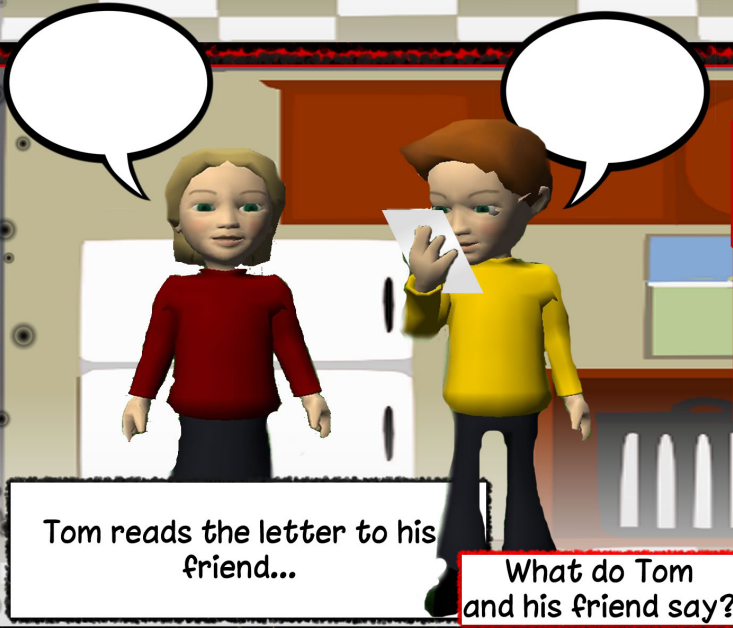
YES NO



THE TRIP

Fill in the missing squares to finish the story.

Tom gets a letter inviting him to go to Summer Camp...



Tom reads the letter to his friend...

What do Tom and his friend say?

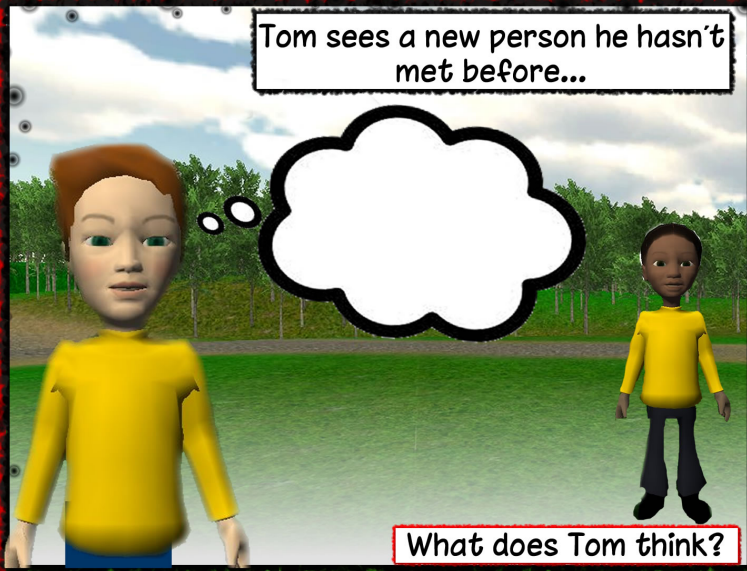
Tom is on the coach to go to the Summer Camp...

Draw a picture to show how Tom feels.

Tom arrives at the Summer Camp...

Draw a picture to show what Tom sees.

Tom sees a new person he hasn't met before...



What does Tom think?

[Empty rectangular box for drawing or writing]

Finish the boxes to show what happens next.

[Empty rectangular box for drawing or writing]

[Empty rectangular box for drawing or writing]

Tom writes a post card to his friend back at home...

[An illustration of an open postcard with blank lined pages for writing.]

THE END

Fill in the post card for Tom.

SUMMER WORDSEARCH

Z S X Y T E A X L V Y T L C W
W P T E A R X A A S H A W L X
O O L L V I A C V G F G U F P
L R S W I M A V G I A R D I B
A T N B R T Q C E N X M H N Q
Z S Q G I R F I J L E C E E V
D D U O S E Y P C S I G J S K
F B N K U M I T M N C J S C W
J T N J Z M M N C A C Z V J A
W R P R F U Q I I K C Y C F T
N G E D I S P B B F E U R P E
J O B L Y Z V N I C S P S X R
J P F L G W U Q P A M D O R J
Z I N S P R X Y B N H R A N M

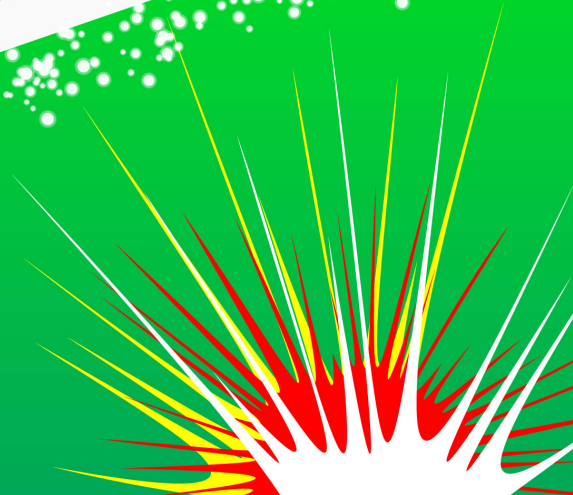
Words to find:

picnic
water
summer

swim
vacation
sports

camp
travel
games

M MAG



HANDS OFF!!

This M-Mag belongs to.....

Today I have.....

.....

.....

.....

.....

Who wins?



Using your sticker pack.
Choose the 3 characters you liked the best
and award them:

1st place for the one you liked the best.
2nd place for the one you liked second best.
3rd place for the one you liked third best.

Place
your
sticker
here



Place
your
sticker
here



Place
your
sticker
here



ROWING REPORTER

The summer camp reporter wants to write a story about Tom.

Answer the questions about Tom below.

How many games of werewolves did Tom play?

1 2 3 4 5

When you first met Tom, did he know how to play werewolves?

YES NO

Who was the best at explaining the rules?
The Yellow or the Red team

YELLOW RED

Would you want to be friends with Tom?

YES NO

Did Tom listen to what you said?

YES NO

Do you think you helped Tom?

YES NO



Was Tom...

Circle one answer from each box below

Good at werewolves

OR

Poor at werewolves

Having fun

OR

Bored

Confused

OR

Knew what he was doing

Good at making new friends

OR

Poor at making new friends

TRUE

OR Use your stickers
to answer if the
question is true or false.

FALSE

Tom thought the yellow team
were cheating?

Tom and his friends were
wearing blue jumpers?

Tom met the yellow team first?



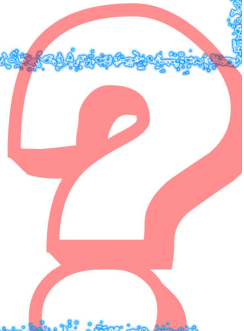
The red team was cheating?

There was a circus down in the MIXER camp?

Tom met the red team first?



In the werewolves game the wolf has to collect easter eggs?



There was a character called Jorden?



MIXER

Tell us what you thought about MIXER by circling an answer for each question

Would you like MIXER to have lasted:

Longer amount
of time

It was just right

Shorter amount
of time

Would you liked to have met another
group of characters in MIXER?

Yes

No

How long do you think MIXER lasted?

5
minutes

10
minutes

15
minutes

30
minutes

Would you want to use MIXER again?

Yes

No

Answer each question by circling
the face that matches your answer the best

**Do you think the game on
the ipad was:**

Easy to use



Completely
agree



Agree



Not
sure



Disagree



Completely
disagree

Difficult to use

Fun



Completely
agree



Agree



Not
sure



Disagree



Completely
disagree

Boring

A good way to
play the game



Completely
agree



Agree



Not
sure



Disagree



Completely
disagree

A silly way to
play the game

**Did you think the pictures
on the iPad:**

Looked
great



Completely
agree



Agree



Not
sure



Disagree



Completely
disagree

Looked
terrible

Were easy
to understand



Completely
agree



Agree



Not
sure



Disagree



Completely
disagree

Were difficult
to understand

What do you think??

Who explained the rules the best?

Tom



or

Yellow
Team



or

Red
Team



Who would you want to be if you played werewolves?



Narrator

or



Werewolf

or



Villager

Which team rules did Tom prefer?

Yellow

or

Red

Do you and your friends play games like werewolves?

Yes

or

No

What did you think about the voices in MIXER?



Liked

Liked a little

Not sure

Disliked a little

Disliked

What did you think about the text in MIXER?



Liked

Liked a little

Not sure

Disliked a little

Disliked

What did you think of MIXER overall?



Liked

Liked a little

Not sure

Disliked a little

Disliked

Did you think MIXER made sense ?



It all made sense

It made some sense

Not sure

It didn't make much sense

It made no sense

In MIXER, did you like the voices or the text the best?

Listening
to voices

or

Reading
text

Friendship Wordsearch

See how many words you can find



Friendship

Friends

Together

Best

Jealous

Argument

Anger

Good

Bracelet

Necklace

Sharing

Nice

Fights

Apologize

Secrets

Trust

Forgive

Promise

Faithful

Everlasting

Polite

Help

Problems

Reliable

ALEX



ANOKI



JILL



JACK



TOM



REKHA



TRACEY



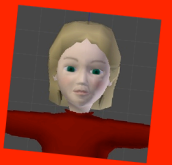
ROB



JORDEN



SUSAN



MARK



True

False

True

False

True

False

True

False

True

False

True

False

True

False

True

False



M=MAG



FINAL EDITION

This m-mag was completed by

.....

NEW PEOPLE, NEW PLACES

Tom has texted you for some advice about making new friends.

Can you help Tom by telling him about you?

Read Tom's messages and dial the number that matches you the best.

- 1 = A lot like you
- 2 = Mostly like you
- 3 = Like you
- 4 = Sometimes like you
- 5 = Not like you



THE EPIC QUIZ!!

ANSWER THE QUIZ QUESTIONS BY TICKING THE BOX THAT IS MOST LIKE YOU

I call people by their real names



A lot like you



Mostly like you



Like you



Sometimes like you



Not like you

I look at people when they are speaking



A lot like you



Mostly like you



Like you



Sometimes like you



Not like you

I walk up to people and start a conversation



A lot like you



Mostly like you



Like you



Sometimes like you



Not like you

I ask questions when talking with others



A lot like you



Mostly like you



Like you



Sometimes like you



Not like you

I look at people when I talk with them



A lot like you



Mostly like you



Like you



Sometimes like you



Not like you

FRIENDS

What kind of friend are you? Circle the face that matches you the most.

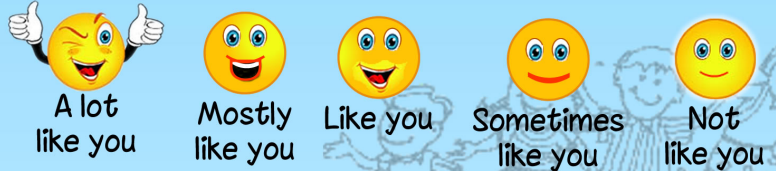
I came to school today with my imaginary friend. When everyone said "hi" to him, I said, "He's just pretend."

But no one seemed to notice, which I thought was pretty weird. It turns out he'd imagined me, and, poof, I disappeared.

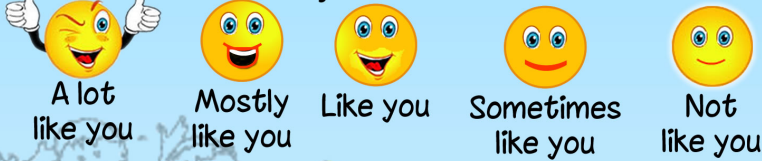
I cheer up a friend who is hurt



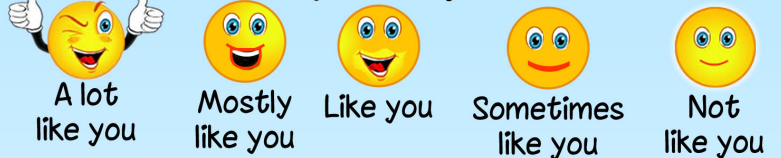
I know how to make friends



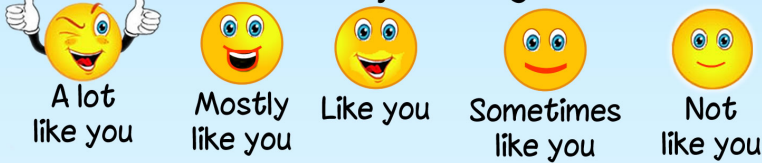
I see my friends often



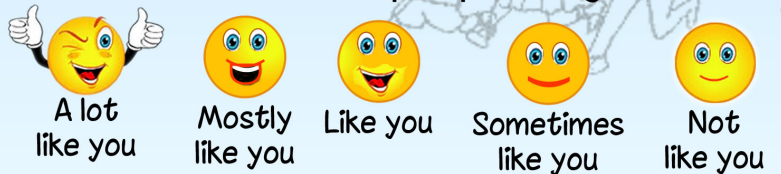
I stick up for my friends



I show my feelings



I make other people laugh

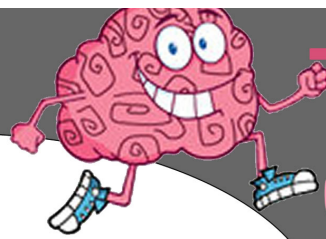


I feel sorry when I hurt someone



THINK FAST

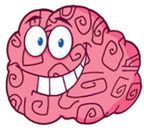
Use your stickers to answer
YES or NO
to the questions below.



Put your
sticker
here

IT'S HARD FOR ME TO
SEE WHY SOMEONE ELSE
GETS UPSET...

I THINK IT'S FUNNY THAT
SOME PEOPLE CRY DURING A
SAD MOVIE OR WHILE
READING A SAD BOOK



Put your
sticker
here

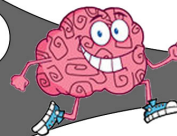
Put your
sticker
here

PEOPLE WHO
KISS AND HUG
IN PUBLIC ARE SILLY!



Put your
sticker
here

IT'S SILLY TO TREAT
CATS AND DOGS AS
THOUGH THEY HAVE FEELINGS
LIKE PEOPLE...



KIDS WHO CRY
BECAUSE THEY ARE
HAPPY ARE SILLY...

Put your
sticker
here

Put your
sticker
here

I DONT FEEL UPSET IF A
CLASSMATE IS PUNISHED
FOR BREAKING THE RULES

I GET MAD WHEN I SEE
A CLASSMATE PRETENDING
TO NEED HELP FROM THE
TEACHER ALL THE TIME...

Put your
sticker
here

Put your
sticker
here

KIDS WHO HAVE NO
FRIENDS PROBABLY DONT
WANT ANY...



I AM ABLE TO EAT
ALL OF MY SWEETS
EVEN WHEN I SEE SOMEONE
LOOKING AT ME
WANTING ONE...

Put your
sticker
here



MAZE DAYS

FIND YOUR WAY FROM THE START TO THE END OF THE MAZE.
ANSWER THE QUESTIONS AS YOU GO.

I laugh at other peoples jokes and funny stories

A lot like you Mostly like you Like you Sometimes like you Not like you

I share what i have with others

A lot like you Mostly like you Like you Sometimes like you Not like you

I join in games with other children

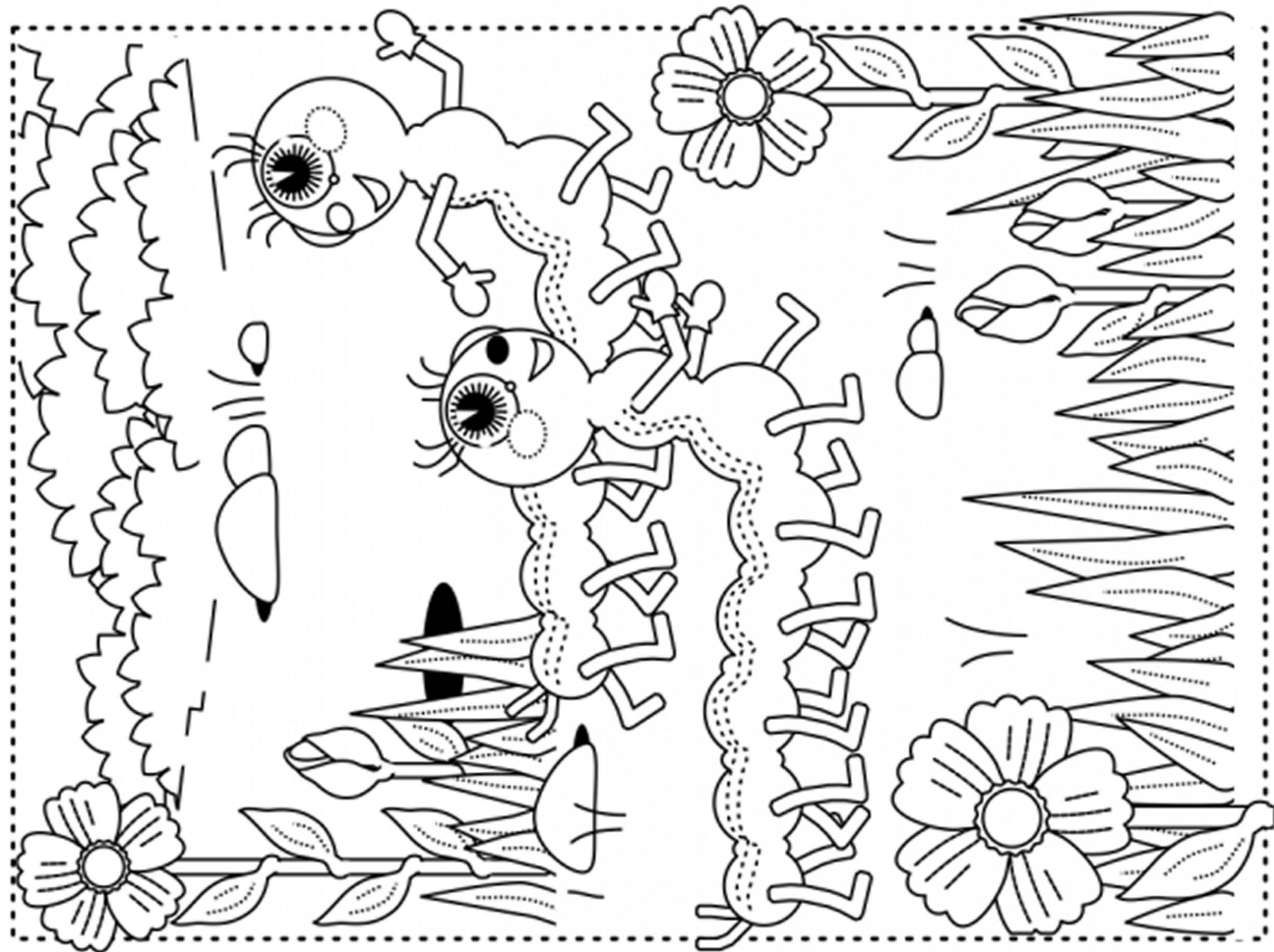
A lot like you Mostly like you Like you Sometimes like you Not like you

START
HERE

END



Find the ten differences between the two pictures.



YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

YES

NO

Five Degrees of Happiness: Effective Smiley Face Likert Scales for Evaluating with Children

Lynne Hall

University of Sunderland
Sunderland, UK

lynne.hall@sunderland.ac.uk

Colette Hume

University of Sunderland
Sunderland, UK

colette.hume@sunderland.ac.uk

Sarah Tazzyman

University of Sunderland
Sunderland, UK

Sarah.tazzyman@sunderland.ac.uk

ABSTRACT

This paper focuses on achieving optimal responses through supporting children's judgements, using Smiley Face Likert scales as a rating scale for quantitative questions in evaluations. It highlights the need to provide appropriate methods for children to communicate judgements, highlighting that the traditional Smiley Face Likert scale does not provide an appropriate method. The paper outlines a range of studies, identifying that to achieve differentiated data and full use of rating scales by children that faces with positive emotions should be used within Smiley Face Likert scales. The proposed rating method, the Five Degrees of Happiness Smiley Face Likert scale, was used in a large-scale summative evaluation of a Serious Game resulting in variance within and between children, with all points of the scale used.

Author Keywords

Question answering; Smiley Face Likert scales; Optimal responses; child-centred evaluation; children

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; H.5.2 User Interfaces (D.2.2, H.1.2, I.3.6) Evaluation Methodology

INTRODUCTION

Typically, most evaluations with children use explicit evaluation activities separate to the interaction (e.g. questionnaires, interviews, panels, etc. [27] and less frequently surveillance techniques (e.g. observation, logging, usage data, etc.). Ólafsson, Livingstone, & Haddon's [24] review of studies of children's use of the internet, identified that over two thirds of studies only collected quantitative data and few studies used mixed methods.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

IDC '16, June 21-24, 2016, Manchester, United Kingdom
© 2016 ACM. ISBN 978-1-4503-4313-8/16/06...\$15.00
DOI: <http://dx.doi.org/10.1145/2930674.2930719>

There are many advantages of using survey methods as they provide a practical and cost effective method of collecting and analysing large amounts of easily anonymisable data. Where available a validated questionnaire will provide a tried and tested method of accurately measuring that, that is to be measured [7,41] improving evaluations and reducing time.

In collecting quantitative data, Tourangeau and Rasinski's [37] 4-stage question response process provides an optimising strategy:

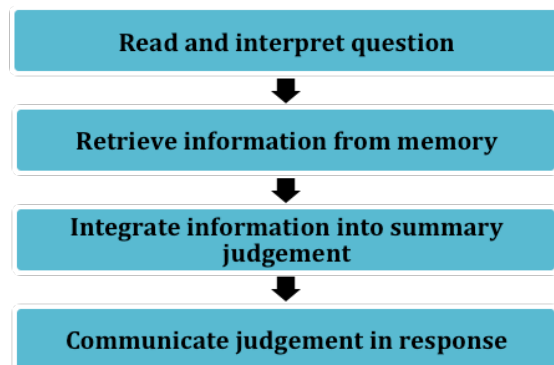


Figure 1: 4 stages of question answering [37]

According to Bell [2], in order for a child to provide an optimal response the following must be true:

1. The child must be able to understand the words and the sentence that forms the question statement
2. The child must be able to associate the question statement with a past experience of their own in order to retrieve the required information to complete step 3
3. The child must understand that the questionnaire is asking them to make a judgment of their past experience against the question statement
4. The child must be able/provided with an effective method to communicate the judgment made in step 3

Whilst all stages merit further investigation, in this paper, we focus on the final stage of this process, an area that has received little consideration. For quantitative questions, the most typical method to communicate judgement is rating scales, with Likert scales a frequently used response item used in evaluation studies with children. Studies have shown that children prefer Likert scales over similar simple

response items such as Visual Analogue Scales [11,16,20]. When used with children, a pictorial Likert scale is often used with images as anchor points. The most commonly used images are smiley faces, which range from negative to neutral to positive, showing very sad to very happy faces, ☹️☺️, [31].

Smiley Face Likerts (SFL) have a long history of use in paediatrics as a subjective measure of children’s medical conditions [36]. More recently SFLs have been used to evaluate children’s opinions of snack preferences [32], of augmented and virtual reality experiences [21,34] and in the use of interactive products [17,22,26,29]. In UX and technological product evaluation with children the use of Smiley Face Likert scales has become common practice, often with aesthetic improvements on the traditional scale as seen in the ‘Smileyometer’ [29].



Figure 2: Smileyometer [29]

However, with children particularly prone to social desirability bias, [28], very positive quantitative evaluations are regularly seen, with the children providing the response that they think the grown-up asking the question wants. Or are they? Could it be instead, that children are not provided with an adequate set of response, thus failing to meet stage 4 of the optimal response process?

Similar to interaction design, evaluation design is fundamentally about engaging users in completing tasks optimally (e.g. answering questions). Yet, there are a lack of papers and practitioner experiences about how evaluations are designed and iterated or evaluations of the evaluations themselves. There is little consideration of whether standard, well-used rating scales do actually provide optimal data, with a wide held assumption that Likerts are fine and SFLs a child-centred way for evaluating children’s experiences effectively.

In this paper, we challenge this view, discussing our investigation into the use of SFLs, gathering data from over 300 children. We highlight the need to change this scale if we really do want a method that allows children to make judgements of their experiences. We discuss how and why we evolved standard SFLs into a tailored, child-centred judgement rating scale. This briefly outlines our progression through a range of studies undertaken in the eCute (www.ecute-project.eu) project using a technology enhanced learning application for 9-11 year olds. Here, unlike most papers on evaluation, we focus on the evaluation process itself, rather than the results generated from that evaluation.

EVALUAND AND EVALUATION CONTEXT

eCute aimed to create and encourage technology enhanced learning experiences to promote cultural awareness, providing intercultural sensitivity learning. It developed MIXER [13], an interactive narrative or Serious Game, aiming to support 9-11 year old children in learning how to recognize and resolve cultural differences. MIXER provides the evaluand for the studies reported in this paper with eCute’s evaluation approach involving multiple formative evaluations feeding into the design of MIXER throughout the lifecycle. In MIXER, see figure 3, the user plays the role of an invisible friend to provide advice and support to a virtual character, called Tom, who is playing Werewolves with a group of virtual characters in a summer camp. Each player is assigned a role, as either a werewolf or a villager. The aim of the game is to deduce which character in the group is the werewolf, before the werewolf kills all of the villagers.

INITIAL DOUBTS ABOUT SFLS

To interact with MIXER, we were developing the Pictorial Interaction Language (PIL) an iPad application with the user dragging and dropping icons to create a dialogue with Tom [8,9]. At an early stage of PIL’s development, we implemented two versions of MIXER for a comparative study between the PIL and a more traditional menu based approach. In Version 1 interaction was via the PIL, in Version 2 the interaction was menu-based providing a set of choices in text form which could be selected by the user by clicking on them (see figure 3 for comparison of the two interfaces).

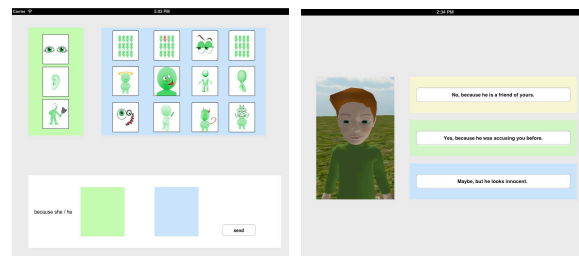


Figure 3: Screenshots of PIL-based interaction versus menu-based interaction

In the procedure, children used each version of MIXER and then completed a questionnaire. Half of the children used Version 1 first and half Version 2 (i.e. the procedure was counterbalanced to avoid order or practice effects). The questionnaire included a series of bi-polar adjectives rated using a 5-point SFL, see figure 4.

An Initial Pilot study with 12 children highlighted a worrying trend... Children tended to rate whatever version they used first very highly, with few negative ratings. Then, when they used the second version even if they found it better than the first they could not rate it higher. However, through observation and child discussions of the two Versions, children clearly preferred the PIL.

Figure 4: Pilot Questionnaire with traditional SFLs

EVOLVING THE SFL: DRAMATIZATION

To increase use of all of the points on the Likert scale we focused on improving the graphical aesthetic of the design. The scale was redesigned to make it more colourful and visual, using cartoon style emojis designed for children. The emotions featured on the smiley faces were dramatized [30], with the intention of evoking a more differentiated approach from children, see figure 5.



Figure 5: Dramatized SFL

To assess the potential of the dramatized SFL, we ran a 29 participant Dramatized SFL study. Children interacted with the PIL and then completed a questionnaire, identical to that of figure 4, except for the change to dramatized SFLs. The results identified an advance in rating variance, with children rating to the third face as well, but no lower. However, as children had been very positive about the PIL, it could be that these ratings were the results of an appropriate method for children to provide judgements. As our focus was to determine which version children preferred, we decided to complement the SFL questions with a question asked at the end of the study (after both questionnaires filled in) where children were given a gold star sticker and asked to put the sticker on a picture of the version they liked the best. Using a simple binary choice such as stickers does have limitations, notably that it does not enable us to know why a child preferred one system over another. However, for eCute, it provided useful evidence to support which interaction approach should be progressed. This use of binary choice stickers should be seen as meeting our pragmatic need rather than as a recommendation towards binary evaluations with children, which yield little information.

The Comparative Study

Seventy one 9-11 year old children participated in a Comparative Study of the two versions of MIXER, with half of children interacting with Version 1 (PIL) first and half interacting with Version 2 (menu-based) first. The questionnaire used the dramatized SFLs and the sticker. Although results from the 8 SFL questions did indicate that in general children rated the PIL version of MIXER higher, there was relatively little difference between the ratings of the two versions.

Children rated all the questions positively for both the menu and PIL interaction, with no mean ratings above 3 (scale ranged from 1 to 5, with 1 being most favourable and 5 the least favourable). All children rated both versions as 3 or higher on all questions. The highest (i.e. least favourable) mean response of 2.61 was for ratings of how exciting / dull the menu-based interaction was.

However, the results from the sticker were much more conclusive, with 92% of children placing their sticker on the PIL version, leaving just 8% (n = 5) of children who placed it on the menu-based version, with an absolutely clear preference. A one-sample sign test revealed that significantly more children said that Version 1 - icon-based was their favourite compared to Version 2 - menu-based [favourite (Z = 6.33, p < .001), words n = 5 (.08), pictures n = 55(.92)].

IDENTIFYING SFLS AS THE CHALLENGE

Study	Children	Variance
Initial Pilot	12	2 or higher
Dramatized SFL	29	3 or higher
Comparative study	71	3 or higher

Table 1: Summary of Early Studies

As detailed in table 1, in using the dramatized SFL, again we were gaining predominantly positive responses, with few children rating 3 and none rating more negatively. In relation to the 4-stage optimal response process, our approach met stages 1-3: our questions had been designed for the age group (e.g. language, developmental aspects); the aesthetic was age appropriate; children's prior experiences (e.g. using MIXER) enabled them to answer the questions. Our study procedure was a traditional, frequently used approach for comparison and counterbalanced to avoid order or practice effects. For stage 4, the children's judgements were provided via the 8 SFL rating and the sticker. With the sticker, 92% of the children identified that the PIL provided a better experience, yet with the SFL, this preference was not clear. This lack of differentiation suggests that we were somehow obtaining sub-optimal responses in response to the 8 rating questions.

Although SFLs are widely used, other researchers have also raised concerns about this rating scale. Zaman, Vanden Abeele, & De Grooff, [39] in their work on comparisons of tangible to other forms of interfaces found the

'Smileyometer' produced results that were inconsistent with children's actual product preferences. Additionally, Mellor & Moore's, [20], recent study on the use of Likert scales with children concluded that children have a limited understanding of the use of Likert response formats. Rubie-Davies & Hattie, [33], also report problems with the use of Likert scales; their results demonstrate that reliability increases with the age of the child but younger children are more likely than older students to respond positively to, and to miss items from Likert scale based questionnaires.

Further, as many studies report, use of such scales can result in straight lining and extreme responding [35]. As with our study, most studies using Likert response formats in questionnaires [4,10,39] [4,10,39] tend to demonstrate extreme positive results, with child respondents agreeing or strongly agreeing to scaled questions. Throughout the literature these results are interpreted as showing that the interactive system is engaging, easy to use, entertaining, etc. Whilst there is some reflection on such positive results, few really ask the question of whether the children's judgements were high quality or sub-optimal. As to why the responses might be sub-optimal, there are a number of biases that can impact on children's judgement and use of such scales in evaluations.

We have already mentioned social desirability bias, where children may not accurately respond regarding socially desirable characteristics in order to appear more appealing to researchers [23]. Specifically for evaluation, this translates to children not wanting to tell an adult that the system they have built is not great. A positive rating is further encouraged through acquiescence bias, or the tendency of respondent's to agree or respond positively [6]. Demand characteristics can also encourage positive responses, with evaluation participants forming an opinion of the purpose of the study and consciously or unconsciously adjusting their opinions or behaviour as a result [19,25]. In all of our studies, we mitigate these biases clearly explaining purpose, highlighting that it is MIXER being evaluated not the children. We strongly emphasize that we are interested in what they really think because we are in a design process.

Less considered, but very important biases for questionnaires include satisficing, a cognitive bias in which respondents decide on and carry out (either consciously or unconsciously) a course of action that will satisfy the minimum requirements necessary to achieve a particular goal. For example, selecting the first reasonable response to avoid reading the rest of the provided options [15]. Satisficing tends to occur if engagement with the evaluation experience is low with respondents seeking the 'path of least resistance' providing a response that satisfies the request made of them by the researcher but which also proves to be the least taxing option for the respondent. Satisficing is seen in straight lining, typically through extreme responding [5]. This bias sees respondents provide

responses at the same, usually extreme, point throughout the scale to either agree or disagree with the statements provided. With children this is particularly common as they tick all the boxes down one side of the page of a questionnaire. In an ideal evaluation respondents would provide an optimal response and therefore one would expect to see variance throughout the responses. A recent finding that held particular resonance for us was that satisficing can also occur because of a lack of differentiation in ratings where scales are provided [38].

EXPLORING SFL SCALE COVERAGE

With concerns about how effective SFLs were in gaining children's judgements, we returned to earlier data, exploring if this lack of variance existed throughout our studies. It did. For example, in [12] we compared 3 sets of questionnaires with identical questions but different look and feel (traditional questionnaire format, questionnaire with limited aesthetic improvement, and a narrative inspired, tailored questionnaire) with 83 children. In both the tailored and the limited aesthetics questionnaire we had used traditional SFLs.

Our focus in this study had been children's engagement with the evaluation instruments, assessed through question completion, abandonment, observed behaviour, questions about the task and time to complete the questionnaires. The tailored questionnaire resulted in complete datasets, no abandonment, no questions and significantly longer time taken to respond to the questions. With our concerns about supporting children's judgements (stage 4) when we returned to the data, we discovered little variance in responses. This was surprising as the questionnaires had not just been user experience but had included personal rating and perception questions from validated questionnaires relating to social skills and cultural awareness.

Whilst we could have rejected the SFL as an inappropriate approach unlikely to generate optimal responses, our results with assessing the engagement with evaluation instruments supported the well-known finding that children had greater engagement with questionnaires that included SFLs. Further, although, means had still been high using the dramatized very positive to very negative SFL, children were prepared to be less positive (e.g. selecting neutral) whilst with the traditional SFL they were only prepared to go as low as the second point on the scale - happy.

Our results highlighted that aesthetically transforming the scale had some impact. However, even with amusing and engaging icons this was not enough to encourage children to use the whole scale. A possible response is to extend the scale and have more categories, however, this increases the complexity of the scale and 5-point SFLs are recommended for children. In response, we began to investigate the research question "What would encourage children to use the full range of available points on an SFL to give appropriate and accurate responses?"

CHANGING FACES

Three iterative studies were undertaken, see table 2, with around 100 children engaging with and assessing MIXER using quantitative questionnaires. For these studies, we were engaging in an iterative design cycle, co-creating and improving PIL’s icons and dialogue structure as well as evaluating the MIXER game as it was being developed, feeding into the design. Each of the studies involved an interaction with MIXER, followed by questionnaire completion. In that, our focus was trying to provide children with 5 points that they might be prepared to select on an SFL scale, we also asked children to rate other activities, e.g. receiving gifts, football and completing homework, with the aim of generating a 5.

With aesthetic and dramatic changes making little difference to using the entire scale we decided to consider the emotions portrayed in the faces. In that no children were rating unhappy and very unhappy we decided to change the SFL. This time the final anchor point was designed to show a face that was only slightly unhappy rather than very unhappy, see figure 6.



Figure 6: SFL with Slightly Unhappy End Anchor

However, none of the 23 children who participated in the Slightly Unhappy Anchor study and completed the questions rated anything, even homework as a negative face, or 5. This suggested to us that children do not want to rate experiences negatively or perhaps, that children consider most things to be at worst neutral and in general positive. This replicates our (and most other evaluator’s) experiences of evaluating very early prototypes where children have been steadfastly positive even if the prototypes have had limited functionality.

In the Neutral Anchor study, we changed the end point of the scale to be neutral, see figure 7, using the questionnaire with 26 children. Again, we incorporated the additional 3 questions, aiming to get a 5. Using the Happy to Neutral scale encouraged four of the children to rate as far as the fifth face, however, this was not for a user experience question, but instead in the rating of homework. Thus, this was an improvement and did suggest we could encourage children to use all of the points on the scale. However, no child rated MIXER lower than a 4.



Figure 7: SFL with neutral anchor point

The results suggest that children do not select negative options, and even when the negative end point was neutral, children were still highly unlikely to select it and not in

relation to evaluating an innovative experience. As to why, well MIXER, like any interactive experience we are evaluating aims to be engaging, entertaining and just generally fun. Thus, perhaps it could be suggested that in the evaluation of interactive experiences, only positive judgements are appropriate. In response, we decided to remove all neutral and negative faces, with the end point changed to a minimally positive face, see figure 8 and conducted a 29 children Slightly Happy Anchor study using both the user experience and additional questions. Use of this scale generated responses across all 5 points, including for ratings of the user experience of MIXER.



Figure 8: The 5 Degrees of Happiness SFL

Our results imply that if we want to provide children with an effective method to communicate the judgment made in response to a question, then the rating scale should provide only positive responses. This scale, the Five Degrees of Happiness, effectively changes SFLs from being a two point rating scale (Positive, Very Positive) to a 5-point rating of what was a positive experience.

Study Name	Children	Variance
Slightly Unhappy Anchor	23	4 or higher
Neutral Anchor	26	4 or higher
Slightly Happy Anchor	29	Entire scale

Table 2: Increasing Happiness of Anchor Studies

USING THE FIVE DEGREES OF HAPPINESS IN THE MIXER EVALUATION

The summative evaluation of MIXER involved a pre-, in- and post- test, with children completing three workbooks, incorporating a range of instruments and activities aiming to assess learning and experience. Workbook One (pre-test) was given to children a week before interacting with MIXER. Workbook Two (in-test) was given to children immediately after their interaction with MIXER. Workbook Three (post-test) a week after the interaction

Workbook One and Workbook Three assessed far transfer of learning. To assess this, Five Degrees of Happiness SFLs were incorporated into the rating scales of the:

- Behavioural subscale of the CQS - Cultural Quotient Scale [1] was used to measure a child’s capability to adapt verbal and nonverbal behaviour in different situations and cultures. In Workbook One (pre-test) the CQS was provided as Woodland Animals and in Workbook Three (post-test) as Maze Days (see figure 9).

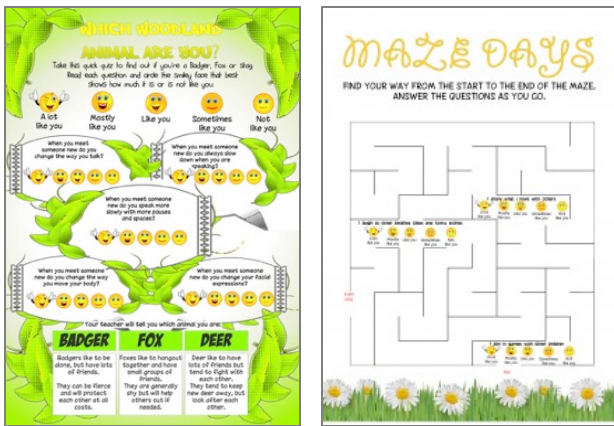


Figure 9: Pre- & post-test CQS

- Factor 2 - Social Skills / Assertiveness of the Matson Evaluation of Social Skills [18] questionnaire used to assess children’s self-perception of their own social skills and competences. In Workbook One (pre-test) MESSY data was collected in New Friendzzz (figure 10). In Workbook Three (post-test) as The Epic Quiz, New People, New Places and Friends (figure 11).

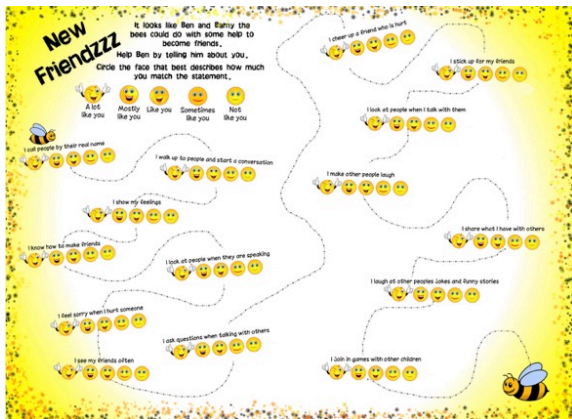


Figure 10: Pre-test MESSY

As can be seen from the figures, particular attempts had been made to make the questionnaires engaging. The designs were inspired by children’s media and co-created with children aiming to create engaging and enjoyable evaluations. In addition, the designs aimed to reduce biases such as satisficing, straight-lining and extreme responding whilst increasing engagement, using age-appropriate gamification and aesthetics. For example, in New Friendzzz (MESSY), the purpose of the activity is to help guide Ben to Barney. The cartoon bees are linked along a dotted line, interspersed with questions. The children move along the line ‘helping’ to get Ben back to Barney and answering the questions as they go. The layout of the questions, which are staggered across the page and follow a curved line, is designed to reduce straight lining. The addition of the line to follow ensures that each question is answered in turn and that no questions are missed out, aiming to create complete

data sets where users are sufficiently engaged in the evaluation to make optimal responses.

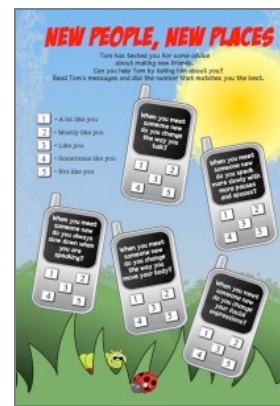
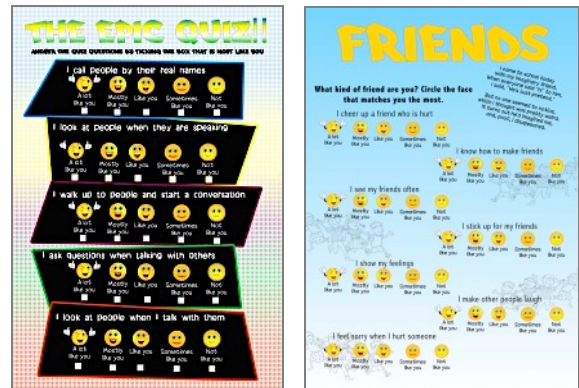


Figure 11: Post-test MESSY

With the workbooks including 30+ questions for children to answer, only some of those that involved Likert scales used the Five Degrees of Happiness. This decision reflects the approach used in activity books for children (e.g. annuals, summer comic specials) where a range of activities and formats are used to maintain interest and the findings of [14] where diversity in instrument aesthetic was identified as critical in not boring users during the evaluation. For example, in ‘New People, New Places,’ an alternative numeric scale is used, see figure 11.

In Workbook Two, the Experience Evaluation Questionnaire, the Five Degrees of Happiness SFL scales were used to evaluate the children’s experience (What do you think?) and evaluate the interaction approach with MIXER (iPad design), see figure 12.



Figure 12: SFLs used in UX questions for MIXER

RESULTS

Over 130 children were engaged in the MIXER summative evaluation, with the results presented in [13]. In this paper our focus is not the evaluation of the evaluand per se, but rather on whether we had managed to have an impact on stage-4 of the optimal response model. Stage-4 requires that children are provided with an effective method that they are able to use and understand enabling them to communicate the judgment made in step 3 (of their experience).

To evaluate whether effective methods had been provided to enable children to communicate a judgement on their experience we used three measures:

- **Completion rates:** this assessed how complete the workbook data were, that is, how many of the rating scales (and thus questions) had the children completed. Low completion rates would indicate a lack of engagement or understanding of the question and rating approach.
- **Individual Variance:** this identified the variance within an individual's responses. High variance (e.g. using the whole scale) would indicate that the SFLs provided children with a method that supported them in making judgements.
- **Sample Variance:** this assessed the variance between participants, determining if within the whole sample the entire scales had been used for each question.

The results are presented in table 3. As can be seen completion rates were almost 100%, with the only incomplete dataset for the CQS in Workbook One (Woodland Animals) where 1 child had not completed this instrument. Sample variance was seen in all workbooks, with all of the scale points selected by at least some children. Individual variance was also high, with the least variance in the CQS in Workbook One (Woodland Animals) and Workbook Three (Maze Days).

Workbook Two provided the in-test measure, the Experience Evaluation Questionnaire. This workbook was

100% complete with 132 respondents. The Five Degrees of Happiness SFL was used for two sets of questions in this workbook. Firstly, relating to the interaction approach, the PIL. Again there was considerable variance, and although most children found the PIL easy, fun and a good way to play with MIXER, we still saw considerable variance, as seen in figure 13.

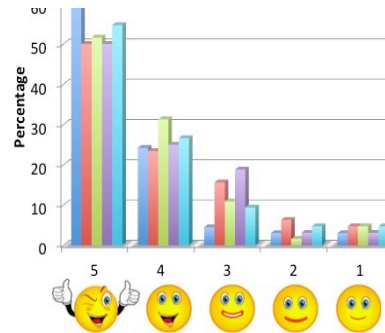


Figure 13: Children's views of the PIL

With the questions relating to the user experience, a good range of variance was seen, for example:

- Children were 'unsure' about the voices in MIXER, median = 3.00, M = 3.23 (SD: 1.44), with scores ranging from 1 = disliked voices to 5 = liked voices
- Children were positive about the text used in MIXER, median = 4.00, M = 3.75, (SD: 1.28), with scores ranging from 1 = disliked text, to 5 = liked text.
- Children felt that MIXER made sense (scale ranging from 1 = 'it made no sense' to 5 = 'it made sense'), M = 4.05 (SD: 1.21), median = 4.00. 11.4% of children said that MIXER 'made no sense' or 'didn't make much sense'.
- Children liked MIXER, M = 4.20 (SD: 1.03), median = 5 (scale ranged from 1 = disliked to 5 liked). 8.3% of children disliked MIXER.

The results from our use of the Five Degrees of Happiness in the MIXER summative evaluation identify that children will use all 5-points of an SFL when the SFL only offers happy emotions.

DISCUSSION

It is often said that childhood is the happiest time of our lives, with news articles claiming that 'children laugh on average 300 times a day compared to adults only laughing 15 times a day.' Whilst this might not be quite true, what is apparent is that children are tuned towards the positive and have a happier mind-set than teenagers and adults. Further, when looking at user experience evaluation of interactive products, we are evaluating experiences that are intended to be fun, interesting and engaging.

	Children	Completion	Individual Variance	Sample Variance
Workbook 1: CQS Woodland Animals	137	136 (99.3%)	127	For all of the questions, there was coverage of all scale points, with at least some children selecting each of the possible SFL scale points.
Workbook 1: MESSY New Friendzzz	137	100% completion	135	
Workbook 2: InteractionPIL Questions (iPad design)	132		129	
Workbook 2: Experience What do you think?	132		130	
Workbook 3: CQS Maze Days	129		113	
Workbook 3: MESSY Epic Quiz, Friends, New P&P	129		127	

Table 3: Results

In all evaluations we have engaged in, children are keen to be entertained. They know that whatever is going to happen is likely to be more fun than a standard lesson. Whilst we have not consistently rated children's 'moods' prior to an evaluation, in the Comparative Study briefly mentioned above, the 71 children were asked to indicate their overall mood before they completed the study. Results ranged from 1 = wow! to 5 = oh dear! using the dramatized SFL. The mean mood rating was 1.67 (SD: .84), illustrating that children were in a really good mood. And every time we have assessed children's mood, they are always in this positive state, expecting to have a great time doing something beyond their usual experience.

If we assume that children are intending to be happy and that we are hoping to give them an interactive experience that is enjoyable, then it is not surprising that children will only select positive ratings. Our early studies identified that children were using 2 points on the traditional SFL, positive, very positive; and we could extend this to the use of 3 points using a dramatized SFL. Whilst however, for children to use the whole scale we had to provide only happy images. Surprisingly this was true both for user experience questions and for self-rating questions (e.g. CQS, MESSY).

A childhood ago, Buckleitner [3] noted, "*As we move into the 21st century, our children deserve rigorous, well constructed evaluation methods applied to the products they use that are subject to public criticism and evaluation.*" However, while researchers are evaluating with children more than ever before, and have increased public availability of results through a significant increase in dissemination and publications there are continuing doubts about the validity of many evaluation results [40]. We had believed that traditional SFLs and aesthetically enhanced variations such as the 'Smileyometer' were effective rating scales, but our results have surprisingly suggested otherwise.

Do anyone else's? We would suggest yes. However, one of the reasons that the evaluation community hasn't challenged SLF results is that they are almost always in our favour. Experience ratings for virtually all interactive products are steadfastly positive when the user group are 9-

11. But as evaluators that is of no help whatsoever, because we need differentiated data.

Our focus on the SLF stemmed from the serendipitous failure of our Initial Pilot study to identify a preferred version of MIXER. This study highlighted that even though the menu-based version of MIXER was lacklustre and very limited, children still had a positive experience.

The series of studies outlined in this paper, identify our evolutionary approach to evaluating SFLs in meeting stage-4 of the optimal response model. There are of course limitations of the research presented in this paper. The approach is practitioner-based, within the context of a live project with a wide range of studies and evaluations typically in the classroom, and represents our consideration and use of SFLs over a 4-year period. For example, the studies comparing increased happiness in the SFL scales were conducted during the lifecycle of MIXER with different children in different classrooms interacting with different scenes, conversations with Tom, etc. in similar although not identical experiences. Thus, the results are not from quite the same experience and we have not attempted to control for such factors. However, as our fairly single-minded aim was to get children to rate something at the negative anchor of the scale, our analysis, prior to the summative evaluation had the single focus: "are any children rating to 5."

With each iteration of the scale, we continued to increase the happiness of the SFLs, certain each time that the scale would generate point coverage. We were surprised to find that to achieve variance, each of the emotions on the SFL needed to be positive. Thus, although intuitively it feels inappropriate to provide no opportunity for children to provide a negative rating (e.g. neutral or unhappy face), in practice perhaps we are imposing an adult answer set that ultimately doesn't provide children with a 5 -point scale.

This approach resulted in the creation of the Five Degrees of Happiness scale that elicited a full range of responses from children. It could be suggested that the problem lies not with the scale but instead is a framing effect. This is unlikely, as the variance in our results indicates that by increasing the happiness of the scale, most children will select across all points.

Our questionnaires are designed to be age-appropriate with appealing, in-narrative inspired aesthetics. We have sought to reduce straight-lining and positive responding using aesthetics and have applied gamification to increase engagement aiming to achieve optimal responses with high variance both between and within subjects. Our use of the Five Degrees of Happiness in the MIXER summative evaluation resulted in complete datasets, very little satisficing and individual and sample variance in use of the scale points. This diversity in the answers suggests that we have managed to provide children with an appropriate method to rate judgments.

CONCLUSIONS

This paper has outlined our exploration of Smiley Face Likert scales for evaluating with 9-11 year olds. Our results highlight that the traditional SFL, with emotions from very happy to very unhappy, has doubtful utility as an effective method for communicating judgments with this age group. This issue is important as we need rating scales methods where children can communicate judgments and that incorporate appropriate differentiation in the scale points. In this paper, we have discussed how we modified and assessed the emotions portrayed in the SFL scale, creating a Five Degrees of Happiness SFL. We have outlined our use of this scale, identifying that it encourages use of all of the scale points, providing an effective method for children to provide judgments in response to scaled quantitative questions.

SELECTION AND PARTICIPATION OF CHILDREN

Over 330 9-11 year old children participated in the studies reported in this paper. The children came from urban state schools in the UK and Germany. Participation included: Initial Pilot 12 children - UK; Dramatized Pilot 29 children - UK; Increasing Happiness of Anchor Studies 78 children - UK and Germany; Comparative Study 71 children - Germany; and the MIXER Summative Evaluation: 137 children - UK. Prior to the study University ethical approval was obtained. Selection was by virtue of them being in the school class that was invited to do the work. Assent and consent forms were provided to the children and parents respectively. The children were told about the aims of the research and when the research was finished they were reminded again and asked if their data could be used. The protocols followed are provided at www.ecute.eu

ACKNOWLEDGEMENTS

This work was partially supported by the EC, funded by the EU FP7 eCute ICT-257666 and FP7 EMOTE ICT-317923 projects. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

REFERENCES

1. Ang, S., Van Dyne, L., Koh, C., Ng, K. Y., Templer, K. J., Tay, C., & Chandrasekar, N. A. Cultural Intelligence: Its Measurement and Effects on Cultural Judgment and Decision Making, Cultural Adaptation and Task Performance. *Management and Organization Review* 3, 3 (2007), 335–371.
2. Bell, A. Designing and testing questionnaires for children. *Journal of Research in Nursing* 12, 5 (2007) 461–469.
3. Buckleitner, W. The State of Children’s Software Evaluation—Yesterday, Today, and in the 21st Century. *Information Technology in Childhood Education Annual* 1999, 1 (1999), 211–220.
4. Chambers, C. T., & Johnston, C. Developmental differences in children’s use of rating scales. *Journal of Pediatric Psychology* 27 (2002) 27–36.
5. Cole, J. S., McCormick, A. C., & Gonyea, R. M. Respondent use of straight-lining as a response strategy in education survey research: Prevalence and implications. *Annual meeting of the American Educational Research Association*, (2012) 1–18.
6. Danner, D., Aichholzer, J., & Rammstedt, B. Acquiescence in personality questionnaires: Relevance, domain specificity and stability. *Journal of Research in Personality* 57, August (2015), 119-130
7. Dowrick, A., Wootten, A., Murphy, D., & A, C. “We used a validated Questionnaire”: What Does This Mean and is it an Accurate Statment in Urologic Research. *Urology*, 85,6 (2014) 1304-10
8. Endrass, B., Hall, L., Hume, C., Tazzyman, S., & Andre, E. A Pictorial Interaction Language for Children to Communicate with Cultural Virtual Characters. In *16th International Conference on Human Interaction* (2014) 532–543).
9. Endrass, B., Hall, L., Hume, C., Tazzyman, S., Andre, E., & Aylett, R. (2014). Engaging with virtual characters using a pictorial interaction language. In *CHI’14 Extended Abstracts on Human Factors in Computing Systems* (pp. 531–534)
10. Guinard, J.X. Sensory and consumer testing with children. *Trends in Food Science and Technology* 11, (2000) 273–283.
11. Haddad, S., King, S., Osmond, P., & Heidari, S. Questionnaire design to determine children’s thermal sensation, preference and acceptability in the classroom. *Proceedings - 28th International PLEA Conference on Sustainable Architecture + Urban Design: Opportunities, Limits and Needs - Towards an Environmentally Responsible Architecture* (2012).
12. Hall, L. and Hume, C. Why Numbers, Invites and Visits are not Enough: Evaluating the User Experience in Social Eco-Systems. *SOTICS 2011, The First*

- International Conference on Social Eco-Informatics*, (2011) 8–13.
13. Hall, L., Tazzyman, S., Hume, C., Endrass, B., Lim, M-Y., Hofstede, G., Paiva, A., Andre, E., Kappas, A. and Aylett, R. Learning to overcome cultural conflict through engaging with intelligent agents in synthetic cultures. *Journal of Artificial Intelligence and Education: Special Issue on Culturally-Aware Educational Technologies* 25, 2 (2015) 291–317.
 14. Hall, L., Jones, S., Aylett, R., Hall, M., Tazzyman, S., Paiva, A., & Humphries, L. Serious Game Evaluation as a Metagame. *Journal of Interactive Technology and Smart Education*. 10, 2 (2013) 130–146.
 15. Krosnick, J. A. The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Newsletter* 20 (2000) 4–8.
 16. van Laerhoven, H., van der Zaag-Loonen, H. J., & Derkx, B. H. F. A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta paediatrica*, 93, 6 (2004) 830–835
 17. Markopoulos, P., Read, J. C., MacFarlane, S., & Höysniemi, J. *Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers*. Morgan Kaufmann Publishers Inc. San Francisco (CA), US. (2008)
 18. Matson, J. L., Neal, D., Fodstad, J. C., Hess, J. a, Mahan, S., & Rivet, T. T. Reliability and validity of the Matson Evaluation of Social Skills with Youngsters. *Behavior modification* 34, 6 (2010) 539–58.
 19. McCambridge, J., De Bruin, M., & Witton, J. The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. *PloS one* 7, 6 (2012) e39116.
 20. Mellor, D., & Moore, K. A. The use of likert scales with children. *Journal of Pediatric Psychology* 39 (2014) 369–379.
 21. Millen, L., Cobb, S., Patel, H., & Glover, T. Collaborative virtual environment for conducting design sessions with students with autism spectrum conditions. *Proc. 9th International Conf. on Disability, Virtual Reality and Assoc. Technologies*, (2012) 269–278.
 22. Nijs, L. and Leman, M. Interactive technologies in the instrumental music classroom: A longitudinal study with the Music Paint Machine. *Computers and Education* 73 (2014) 40–59.
 23. Oerke, B. and Bogner, F.X. Social Desirability, Environmental Attitudes, and General Ecological Behaviour in Children. *International Journal of Science Education* 35, 5 (2013) 713–730.
 24. Ólafsson, K., Livingstone, S., & Haddon, L. Children's use of online technologies in Europe: a review of the European evidence base. EU Kids Online, London, UK (2013)
 25. Orne, M.T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist* 17, 11 (1962) 776.
 26. Read, J. MESS Days: Working with Children to Design and Deliver Worthwhile Mobile Experiences. *UPA User Experience Magazine*, 9, 2 (2010)
 27. Read, J. and Markopoulos, P. Evaluating children's interactive products. *Extended abstracts of the 32nd annual ACM Conference on Human Factors in Computing Systems*, (2014)1043–1044.
 28. Read, J. and Fine, K. Using survey methods for design and evaluation in child computer interaction. *Workshop on Child Computer Interaction: Methodological Research at Interact*. (2005).
 29. Read, J., MacFarlane, S., & Casey, C. Endurability, engagement and expectations: Measuring children's fun. *Proceedings of Interaction Design and Children*, (2002) 189–198.
 30. Reynolds-Keefer, L., Johnson, R., & Carolina, S. Is a picture worth a thousand words? Creating effective questionnaires with pictures. *Practical Assessment, Research & Evaluation* 16 (2011) 1–7.
 31. Reynolds-Keefer, L., Johnson, R., Dickenson, T., & McFadden, L. Validity issues in the use of pictorial Likert scales. *Studies in Learning, Evaluation Innovation and Development* 6 (2009) 15–24.
 32. Roberto, C. A., Baik, J., Harris, J. L., & Brownell, K. D. Influence of licensed characters on children's taste and snack preferences. *Pediatrics* 126 (2010) 88–93.
 33. Rubie-Davies, C. M., & Hattie, J. A. C. The dangers of extreme positive responses in Likert scales administered to young children. *The International Journal of Educational and Psychological Assessment* 11 (2012) 75–89.
 34. Salvador-Herranz, G., Perez-Lopez, D., Ortega, M., Soto, E., Alcaliz, M., & Contero, M. Manipulating virtual objects with your hands: A case study on applying desktop Augmented Reality at the primary school. *Proceedings of the Annual Hawaii International Conference on System Sciences* (2013) 31–39.
 35. Sluis, F. Van Der, Dijk, E. M. a G. Van, & Perloy, L. M. Measuring Fun and Enjoyment of Children in a Museum : Evaluating the Smileyometer Study One : Prototype. In *Proceeding of Measuring* (2012) 86–89.
 36. Tatla. 2014, S.K. The development of the Pediatric Motivation Scale for children in rehabilitation: a pilot study. Retrieved from http://elk.library.ubc.ca/bitstream/handle/2429/45920/ubc_2014_spring_tatla_sandeep.pdf?sequence=27
 37. Tourangeau, R. and Rasinkski, K.A. Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin* 103 (2008) 299–

314.

38. Vannette, D. and Krosnick, J. A comparison of Survey Satisficing and Mindlessness. In *The Willey Blackwell Handbook of Mindfulness*. (2014) 312.
39. Zaman, B., Vanden Abeele, V., & De Grooff, D. Measuring product liking in preschool children: An evaluation of the Smileyometer and This or That methods. *International Journal of Child-Computer Interaction* 1 (2013) 61–70.
40. Zaman, B., Vanden Abeele, V., Markopoulos, P., & Marshall, P. Editorial: The evolving field of tangible interaction for children: The challenge of empirical validation. *Personal and Ubiquitous Computing* 16, (2012) 367–378.
41. Zarins, B. Are validated questionnaires valid? *The Journal of Bone & Joint Surgery* 87, 8 (2005) 1671–1672.

Engaging Children In Interactive Application Evaluation

Lynne Hall, Colette Hume
University of Sunderland

Sarah Tazzyman
University of Leicester

Author Note

Lynne Hall and Colette Hume, Department of Computing, Engineering and Technology, David Goldman Informatics Centre, University of Sunderland, St Michael's Way, Sunderland, SR6 0DD. Email: lynne.hall@sunderland.ac.uk, me@colettehume.com

Sarah Tazzyman, Henry Wellcome Building, The School of Psychology, University of Leicester, Leicester, LE1 9HN, UK. Email: snt5@le.ac.uk

Correspondence concerning this article should be addressed to Lynne Hall, Department of Computing, Engineering and Technology, David Goldman Informatics Centre, University of Sunderland, St Michael's Way, Sunderland, SR6 0DD. Email: lynne.hall@sunderland.ac.uk

Abstract

Interactive applications designed specifically for children offer great potential for education and play. However, to ascertain that the aims of applications are achieved, child-centred evaluations must be conducted. The design of any evaluation with children requires significant consideration of potential problems with comprehension, cognitive ability, response biases and study attrition. Multidisciplinary R&D project evaluation requirements are often extensive, requiring an all-encompassing and prolonged evaluation design. Discontinuity between the highly engaging interaction experience and the multitude of measures that form the evaluation poses a major issue for the evaluation of interactive applications. In response, we have developed Transmedia Evaluation, a method that aims to maintain engagement throughout the evaluation process. In this paper, the Transmedia Evaluation process is explained and applied to evaluate a learning application for children, MIXER (Moderating Interactions for Cross Cultural Empathic Relationships). Children aged 9-11 (N=117) used the MIXER application and completed an evaluation battery including pre- and post- test questionnaires, immediate learning assessment and qualitative evaluation. Using Transmedia Evaluation to develop the MIXER evaluation resulted in complete data-sets (100%) for quantitative data (by self-regulated completion) along with rich, high quality qualitative responses. Transmedia Evaluation transformed the evaluation, with children fully engaging in and enjoying their experience.

Keywords: evaluation, child-centred design, user experience, learning technology

1 Introduction

In evaluating children's experience of interactive applications we, as researchers and evaluators, are aiming to provide further evidence for or against specific issues, expectations and concerns related to the impact of the interaction on the child. Whilst innovations and experiments across the reality spectrum have produced a myriad of engaging applications for children, this trend has not been followed in their evaluation. Although there has been a significant increase in studies about children's use of interactive technologies, this hasn't resulted in a significant diversity of methods used to gather evaluation data. With rare exceptions, the evaluation of even the most radical system has relied on surveillance techniques (e.g. video observation, logging, usage data, etc.) and/or explicit evaluation activities (e.g. paper/pencil questionnaires, interviews, panels, etc.). In Ólafsson, Livingstone, & Haddon's, (2013), review of studies of children's use of the internet over two thirds of studies only collected quantitative data and few studies used mixed methods.

Interactive applications developed for children often intend to immerse and engage them within a self-created and maintained experience. Yet, when the focus turns from interaction to evaluation, this immersion is frequently fractured. The focus, design, specific tasks and overall image of evaluations are

ENGAGING CHILDREN IN INTERACTIVE APPLICATION EVALUATION

often significantly different, and at odds with the interactive experience. Whether for games, recreation, learning or social environments, evaluation is often disruptive, provided as a separate, dislocated activity, see figure 1, with little consideration of the user's experience. For children, this can result in being taken from being engaged and having fun in roles such as 'virtual pet owner' 'secret friend' or 'space cadet' to instead being placed into the role of 'subject' in an evaluation procedure. A standard-format questionnaire can be viewed as a disengaging follow-up activity, especially if it follows a novel and immersive technological experience.

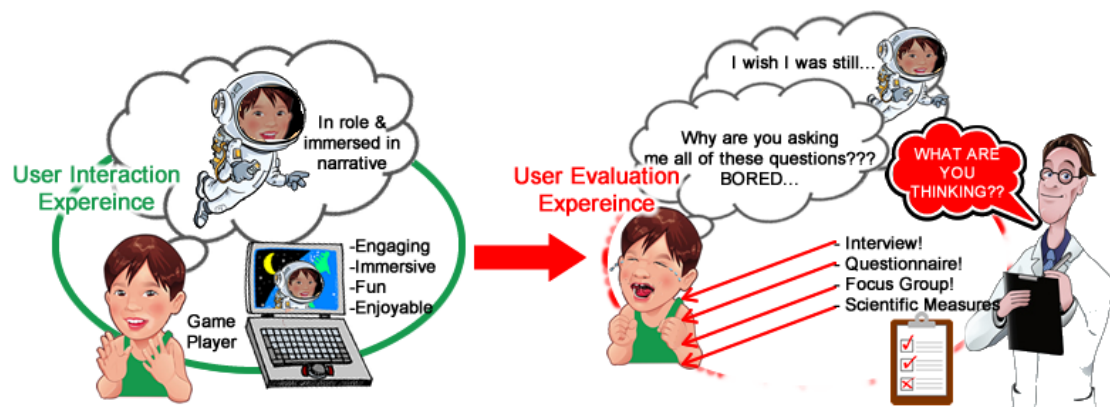


Figure 1: Standard 'disruptive' evaluation approach

In R&D evaluations of children's use of technology the primary instrument is questionnaires. Administration is typically straightforward and data analysis from structured questionnaires provides a well understood and accepted evaluation methodology throughout the research, public sector and business communities. However, child-centred factors that can impact on question answering, such as developmental effects including language ability, reading age, and motor skills, as well as temperamental effects such as confidence, self-belief and the desire to please (Read & MacFarlane, 2006) are rarely dealt with in the evaluation design. Many evaluations involve children filling in instruments that use adult language and formats, continuing the trend noted in Jensen & Skov, (2005). Although some evaluations do attempt to create appropriate methods, in general, most evaluations for children are very similar to adult evaluations, where interaction is surrounded by arduous, possibly unappealing and frequently inappropriate evaluation instruments. This can all result in study attrition and incomplete data sets, which can greatly impact on the overall results and conclusions drawn from the evaluation.

Using traditional evaluation approaches with children can have serious implications, both for the child's experience and the quality of data collected. A lack of engagement typically results in providing sub-optimal responses in questionnaires, with a high chance of satisficing (Krosnick, Narayan, & Smith, 1996) and acquiescence bias (Babbitt, 1989). Usability and user experience satisfaction studies tend to demonstrate extremely positive results, with child respondents agreeing or strongly agreeing to scaled questions. Throughout the literature these results are interpreted as showing that the interactive system is

engaging, easy to use, entertaining, etc. Few really ask the question of whether the data was high quality or sub-optimal. This can have important implications for conclusions drawn and future development. For instance, Buckleitner, (1999) noted *“As we move into the 21st century, our children deserve rigorous, well constructed evaluation methods applied to the products they use that are subject to public criticism and evaluation.”* However, even though researchers are evaluating with children more than ever before, and have increased public availability of results through a significant increase in dissemination and publications there are continuing doubts about the validity of many evaluation results (Zaman, Vanden Abeele, Markopoulos, & Marshall, 2012)

Child representation and respect are further issues raised in the evaluation of interactive applications for children, highlighted by Read et al., (2008) who note that *“A core value for the field of Child-Computer Interaction is that the interests of children are represented and respected in the research and design processes.”* However, in many evaluations there appears to be very little representation of or respect for children’s interests Nor do studies typically report on children’s response to evaluation, although Sapouna et al., (2010) note that the additional activity required by evaluation can diminish the child’s enjoyment of the experience. With the focus of evaluation on the capture of valid and reliable data to substantiate hypotheses, the centre of an evaluation design is not the child, but rather the R&D motivation. Appropriately designed evaluations need to place children at the centre of the evaluation experience, just as we recognize that we should place them at the centre of the interaction design.

This paper discusses Transmedia Evaluation, a methodology for creating evaluation experiences that places users at the centre of the design. The approach aims to seamlessly embed evaluation into the user experience, providing valid and reliable data and adding value to the user. Transmedia Evaluation was developed and trialed with 9-11 year old children as the primary users and critical participants in the evaluation. We focus on the Transmedia Evaluation of MIXER, a technology enhanced learning application targeting intercultural conflict developed for 9-11 year olds (eCute, 2012), that provides users with immersive virtual role-play with intelligent interactive graphical characters.

2 The Evaluand: MIXER

MIXER, see figures 2, 3 and 4, is a Virtual Learning Environment populated by intelligent, affective and interactive characters targeted at 9-11 year old children, highlighting strategies and supporting the development of intercultural skills and competences. The Summative Evaluation of MIXER aimed to provide demonstrable evidence that experiential intercultural learning could be provided to children through the innovative technology, further detailed in (Aylett et al., 2014; Endrass, Hall, Hume, Tazzyman, & Andre, 2014), developed in the eCute project

MIXER engages users in an interactive narrative set in a virtual summer camp, where two groups of school children (intelligent characters) play Werewolves, a popular intergenerational game widely known in many cultures. As is common in summer camps, the children were dressed in team T-shirts, see figure 2, representing the two teams: the Reds and the Yellows.

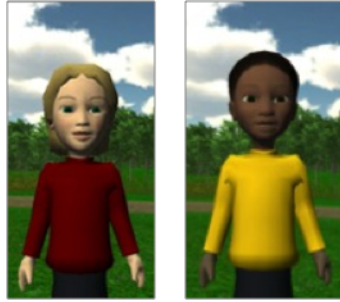


Figure 2: Alex and Lisa, characters from MIXER

MIXER depicts a peer conflict scenario, occurring when Tom (protagonist) plays the Werewolves game with two different teams of children, the Yellows and the Reds at a summer camp, see figure 3. Each team plays by a different rule set resulting in a conflict situation for Tom who subsequently accuses the red team of cheating, because he does not understand the rule change. MIXER ends by Tom resolving the conflict with the Red team by discussing the differences in the two versions of Werewolves.



Figure 3: Scenes from MIXER

In MIXER, the child does not directly appear in the virtual world. Instead their role is to interact with Tom, as an invisible friend and to support his play by responding to Tom's requests for advice on how to react and what to do at different stages of the game. The child interacts with Tom through a tablet using a Pictorial Interaction Language (Endrass, Hall, Hume, Tazzyman, Andre, et al., 2014), (see figure 4), providing children with access to over 70 graphics structured for use in sentences, enabling them to interact with Tom.

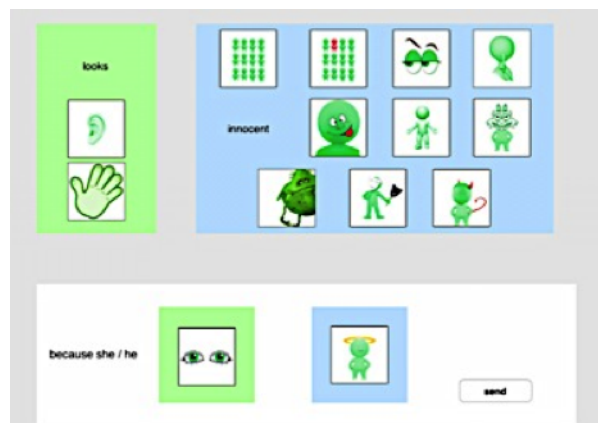


Figure 4: Fragment of Pictorial Interaction Language

3 Transmedia Evaluation: Background & Basis

Talking to children about evaluation quickly identifies that their expectations are constructive and optimistic. Children expect to have an interesting, entertaining, and engaging experience, whatever it is they are expected to do. Placing this expectation of enjoyment and engagement on evaluation, quickly changes the nature of the activity, away from the traditional approach of 'doing something to someone to gather data for R&D purposes' instead 'to designing an engaging experience for the user enabling them to provide quality data.'

Research has rarely considered creating engaging evaluation experiences of interactive applications, whilst there has been considerable focus on enhancing engagement. Engagement is viewed as a quality of user experience that facilitates more enriching interactions with interactive applications (O'Brien & MacLean, 2009; O'Brien & Toms, 2010). Further, it can be defined by a core set of attributes: aesthetic appeal, novelty, involvement, focused attention, perceived usability, and endurability. Designing and implementing these attributes into evaluation experiences would clearly create more engaging and enriched experiences. Whilst it is relatively straightforward to create usable (e.g. sensible number of age appropriate questions) and appealing (e.g. age appropriate graphics) materials, incorporating attributes such as involvement and focused attention is more challenging.

Engaging users requires a dramatic rethink of how we present the experience to the user. Our approach has been inspired by transmedia: *"...a process where integral elements of a fiction get dispersed systematically across multiple delivery channels for the purpose of creating a unified and coordinated entertainment experience"* (Jenkins, 2011). The most successful transmedia encircles and extends the primary user experience (e.g. viewing a movie or programme or in our case, engaging with an interactive application), taking the narrative from a TV show or movie to create a nucleus that is surrounded by supplemental story lines and activities. Transmedia is "a user-focused experience that is collaborative, immersive, and interactive" (Parker & McDonald, 2014). In contrast to evaluation, the additional experience and activity offered by transmedia adds considerable value to the user experience and users wish to engage with it.

Transmedia, content must be compatible with the themes, tone and message of the film (Gomez & Pulman, 2012), authentically extending the story world in which the experience unfolds (Weiler, 2012). As such, Transmedia Evaluation aims to provide users with a unified, themed and coordinated experience providing consistent, integrated content through appropriate channels, platforms, devices and activities designed to meet user expectations and to reinforce engagement with the experience.

As figure 5 depicts, Transmedia Evaluation aims to seamlessly embed evaluation into the user experience by creating evaluation materials and activities that are both appropriate and engaging for the target user group, connect to the evaluand and result in high quality data for the research team. In large R&D projects, these goals cannot be met by simply embedding the evaluation within the evaluand. Rather Transmedia Evaluation aims to integrate interaction with an innovative,

interactive system and the related evaluation battery into a consistent, coherent user experience.

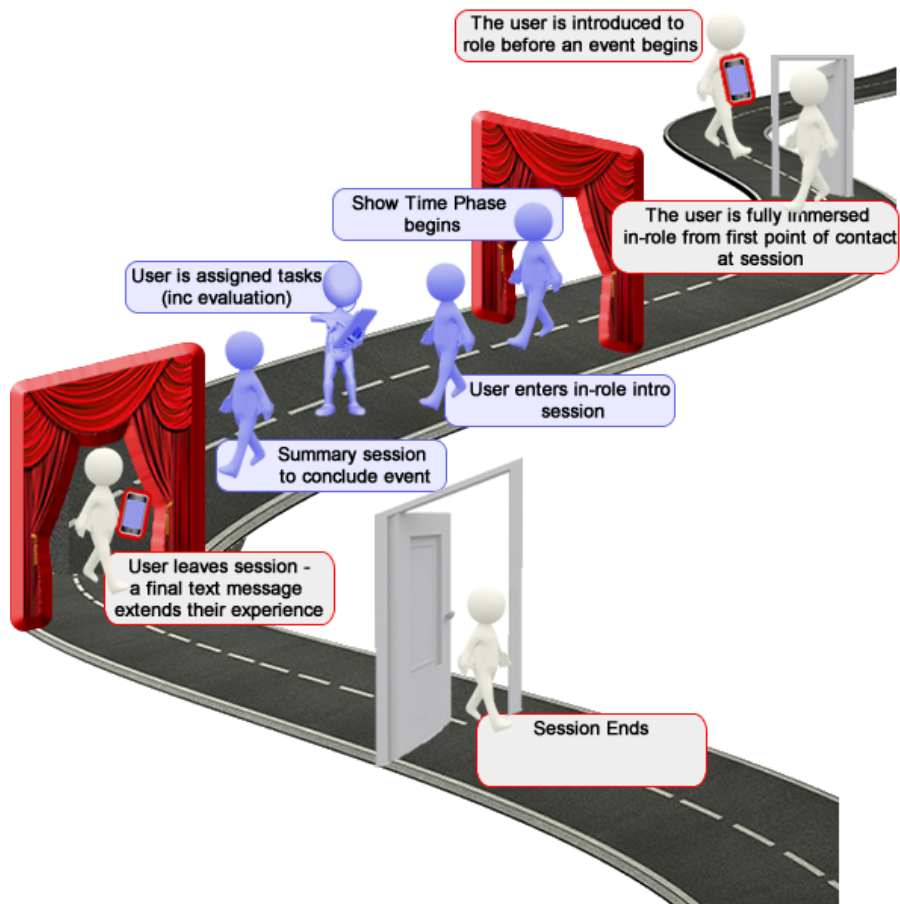


Figure 5: A Transmedia Evaluation Event

Early piloting of Transmedia Evaluation focused on ensuring that R&D requirements were met even if instruments had been transformed to provide children with a single, coherent, transmedia inspired experience. Using a low fidelity evaluand of MIXER (a comic strip), three variants of the same instruments were provided (see figure 6):

- **Basic**, traditional evaluation approach (A4, black and white, numbered quantitative and qualitative questions – age appropriate language and format);
- **Better**, more hybrid approach, providing cosmetically improved instruments for example appealing colour graphics, interactive activities and some variety in question and response formats, but without a clear connection to the evaluand
- **Best**, as an integrated comic book incorporating the MIXER comic strip with evaluation materials, based on the cosmetically improved instruments but designed to reinforce a connection with the evaluand (e.g. using figures from the comic strip in evaluation activities).

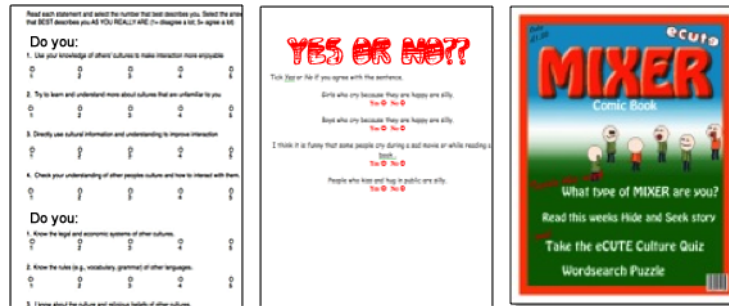


Figure 6: Basic, Better and Best materials from early stage evaluation

The results were startling. Not only was appropriate data provided even in the most transformed of the instruments, but further this data was more complete, of better quality, showed richer qualitative responses and improved user engagement (Hall & Hume, 2011). The more the evaluation materials met the children's expectations (e.g. the better they looked, the more interactive they were and the more they connected to the evaluand), the more the children engaged and the higher the quality of data. These initial positive results inspired the development of an engaging methodology designed to be value laden for the user.

4 The Transmedia Evaluation methodology and its application to the MIXER Summative evaluation

Figure 7 provides an outline of the Transmedia Evaluation methodology. This supports the development of an evaluation providing the plot (R&D perspective); role (intended user experience); props (evaluation battery and evaluand); and the script (experience protocol) required for a Transmedia Evaluation event. The event, and the elements within it, are rehearsed and refined (piloted, incrementally iterated), with the performance of the event (all aspects of user experience including evaluation, training (if required) and interaction) followed by a review phase (evaluation of event and data), which then feeds back into subsequent evaluations. The following sections further detail each of these stages, outlining how the approach was applied to the Summative Evaluation of MIXER.

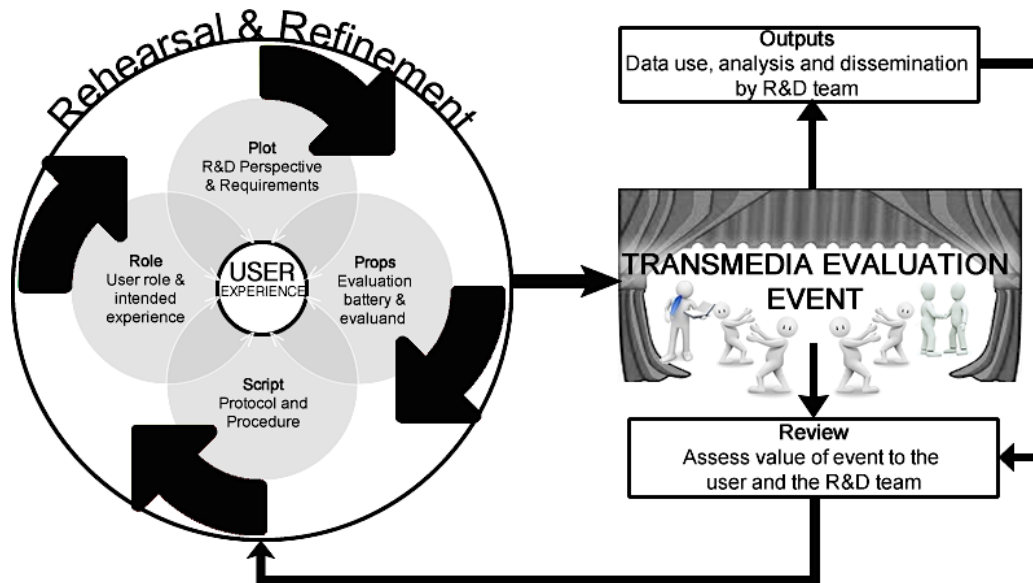


Figure 7: Transmedia Evaluation Framework

4.1 The Plot - R&D Perspective & Requirements

Transmedia Evaluation begins with an outline 'plot' providing specific R&D hypotheses, constraints (e.g. setting, participant numbers, interactions with evaluand, training requirements) and empirical parameters (e.g. within group, between group; qualitative, quantitative). The R&D requirements provide the key elements that must be incorporated into the user experience to achieve an effective evaluation from the perspective of the R&D team.

The plot for the MIXER Summative Evaluation was to identify if the learning goals of the interaction as specified in the eCute Intercultural Competence Learning Framework (Swiderska, Krumhuber, Kappas, Degens, & Hofstede, 2011) were met:

- Emotional: MIXER supports children to recognise emotions (for example fear and anxiety) when dealing with the strange behaviours of another group
- Cognitive: MIXER supports children to start learning the specific practices and values of another group
- Behavioural: MIXER supports children in being fully present in attending to others verbal and non-verbal messages

A further goal was to determine whether the MIXER technology (e.g. intelligent agents, interaction modality, emergent narrative) was an effective approach for technology enhanced learning:

- Experience: MIXER engages children in the narrative and with the characters, supporting the children's understanding and learning of strategies for coping with intercultural conflict

The evaluation of MIXER's impact on children's learning provided the R&D requirements of a controlled randomized pre- post- design, collecting quantitative data to enable the assessment of far transfer (e.g. sustained learning). Within the test, R&D requirements identified the need for evaluation

to include qualitative and quantitative measures to assess near transfer (e.g. immediate learning); and the user's response to the underpinning technology as provided by the characters and the interaction modality. Subsequent to the interaction, a reflective session to reinforce children's learning of intercultural conflict had to be incorporated into the evaluation design. The evaluation was designed to be classroom-based, involving 100+ children.

The plot for a Transmedia Evaluation Event is developed as a series of nodes or acts, within which users have to perform certain activities (such as interacting with the evaluand) or certain elements of the evaluation (e.g. participant information questionnaire). Plot development requires the evaluators to identify established instruments, data capture approaches and activities that can be used to assess and meet R&D hypotheses. Transmedia Evaluation advocates the use and/or adaption of existing measures and techniques wherever possible as this improves the reliability and validity of the data. Obviously, there are contexts where no measure or activity exists, for example, in assessing evaluand specific hypotheses (e.g. assessing a user's comprehension of specific story elements in a storytelling application).

In the MIXER Summative Evaluation plot, the three measures aimed at assessing far transfer according to the specified learning goals were taken from the CQS - Cultural Quotient Scale (Ang et al., 2007); the MESSY Scale - Matson Evaluation of Social Skills (Matson et al., 2010) and Bryant's Empathy Index (Bryant, 1982). The behavioural subscale of the CQS was used as a pre- and post- measure of a child's capability to adapt verbal and nonverbal behaviour in different situations and cultures. Factor One from the Bryant Empathy Index focuses on understanding feelings and was used as a measure of children's empathic behaviour. Factor 2 - Social Skills/Assertiveness of the MESSY questionnaire was used to assess the child's self-perception of their own social skills and competences.

Quantitative measures to assess the user's engagement, interaction and immediate learning were based on questionnaires developed for assessing the user experience in VLEs populated by embodied characters, based on Hall et al., (2013) and Hall, Woods and Aylett, (2006). Theory of Mind questions required by the R&D team used to assess children's advice to Tom are embedded in the conversation the child has with Tom, following the approach in (Hall, Woods, & Hall, 2009). For example in advising Tom during the conflict incident, Tom asks the child what he should do, why, what makes the child think that will work, etc.

As the evaluation event is piloted (Rehearsal) and further developed, the plot is extended to incorporate specific instruments, activities and data capture approaches. Finalised instruments are supported with relevant protocols, merging them into the script. Coding frames and datasets are provided ensuring that analysis can begin rapidly after the evaluation event has finished.

The plot aims to provide a structure that is infused by the user role, by creating coherent and engaging props and scripts, see figure 8. The interaction with the technology and evaluation becomes one part of a connected, coherent experience for the user who is having a great time using new technology and participating in engaging activities.



Figure 8: MIXER Plot for Summative Evaluation

4.1.1 Role – User Role & Intended Experience

The user is at the centre of a Transmedia Evaluation event, with the experience designed to meet the most basic user expectation of having an enjoyable time. Initial considerations of the user typically involve a review of current literature, applications, media and on-/off-line activities and experiences, using techniques frequently seen in persona creation. This exploration of the user's world aims to immerse the evaluators, inspiring and informing them about what interests and engages the target users.

The user role must be sympathetic to the evaluand, connecting with this in a way that is consistent, comprehensible and credible for the user. The user role may be an in-application role such as playing a character in a specified storyworld setting. It can also be in-task roles of learner, storyteller, player, etc.; traditional evaluation roles such as subject, designer or critic; and even that of the user's everyday self, effectively not changing role at all. Transmedia Evaluation places the user quickly in role for their experience, with recruitment reinforcing the user expectation of having a good time both by introducing their in-experience role and by highlighting their value to us. Recruitment must not only achieve informed consent, but must also reinforce the sense that the children are going to participate in something interesting, novel, important and relevant to them.

Typically for any evaluation, only a limited number of roles are possible. For instance, with a game if the user role was as player, then the evaluation experience and artefacts must become part of the game, by expanding the experience of the game world such as completing a self-rating scale as part of the entry requirements to a guild. If the user role was as critic, then the experience and artefacts must support the user in critical activity in a way that meets their expectations, for example completing rating scales (e.g. how many stars the game merits) or posting reviews to a Critic's Website.

The user role unifies the various elements of the experience, just as transmedia is unified by the overarching theme of the film or programme that it encircles.

Ideas for user role, along with initial props, such as instruments and early versions of the evaluand, are piloted with the target user group gaining their input.

A range of user roles were considered for the users of MIXER, including:

- Related to the user's role in MIXER (invisible friend) with the child being Tom's friend in the evaluation experience
- Related to MIXER's storyworld (but not in the interaction) with the child's role being as camp counselor, for example.
- Related to MIXER's aim with the child's role being as a learner

Maintaining in-application user roles throughout the pre- and post- test phases, and particularly in incorporating the repeated 3 measures for far transfer and the learning reinforcement experience highlighted that fracturing of in-application roles was likely, thus rejecting placing the user in the role of Tom's friend. Although placing the user in a role such as camp counselor was considered, we decided against this as it implied that the user was operating at an expert level (e.g. already able to help and advise others) rather than as a novice learner. With the need to fit within the school day and to engage over multiple, separate sessions with the children, we refined the user role as 'learner' to the children's everyday role as a school-based learner, with the evaluation event being one of the children's lessons within the school day.

A review of information about children's interests, expectations and activities engaged in, informed the user role, with the aim not just for the child to have an average lesson, but rather a user role where the child is having an excellent experience using state-of-the-art technology to learn something different. Our interpretation of this role can be seen in the props detailed below.

4.1.2 Props (Evaluation battery & evaluand)

All evaluation instruments and approaches that are visible and require active participation by the user (as opposed to surveillance, covert data collection.) are viewed as props, integrated into the plot and user role. Transforming the instruments into props is an incremental, iterative process and users are involved in the design and piloting of all evaluation props.

Each of the instruments and evaluation points identified in the plot is initially provided in a basic form. For example, with questionnaires the usual approach to administering the instrument is provided, this is often black and white, with numbered questions often with Likert rating scales or categories. Qualitative issues and questions (e.g. for interviews and focus groups) are listed. Techniques are provided as brief outlines, indicating required activities or outputs. At the start of the transformation process, a key issue to be addressed is "how appropriate is the intended instrument and battery for the intended user group?" With many questionnaires incorporating multiple sub-scales or factors and possible duplication between proposed instruments, the initial transformation ensures only necessary data is collected.

Once the instruments are provided, instrument refinement then permits further assessment and improvements if deemed necessary. Irrespective of aesthetic appeal, if questions don't make sense, seem repetitive or burdensome to answer, then user responses will be less optimal. As only representative users can tell

you if the questions are appropriate, this transformation requires piloting with users.

An immediate issue in using the identified far transfer measures for the MIXER Summative Evaluation was the number of questions (with 104 questions across all three questionnaires) and the adult focus. Incrementally, with the R&D team, the instruments were refined, for example only using the behavioural subscale of the CQS. Sessions with users were held to improve the language and comprehensibility of the measures.

With the basic prop confirmed, the second level of transformation aims to reinforce the user role and to connect the prop to the evaluand. Qualitative data collection readily lends itself to the reinforcement of user role and integration into the plot enabling the collection of required data. In many studies, qualitative data is collected as written or spoken answers to open questions, with considerable flexibility as to how these questions are asked. There are many natural ways to incorporate such data collection into almost any user role and age group, using text (e.g. postcards, notebooks, posters), verbal (e.g. focus groups, interviews) and digital (e.g. texts (SMS), selfies, user-generated media) approaches. For example, if we put the user into the role of a 1900's news reporter with a history focused evaluand and then ask them to provide short, qualitative data about their learning (e.g. story comprehension, fact identification) via mobile phone to call a friend, the user's immersion with this out of place prop and reference to modern TV games shows would be ruptured. A more fitting prop would be a notebook into which the user could make notes on the events around them and then post these to an editor. Although the user sees nothing more than stage props in the story world experience, these items are actually the transformed evaluation materials that collect qualitative data and reinforce the user's role as in experiencing the 19th century context of the evaluand.

In the MIXER Summative Evaluation qualitative data was collected both from the qualitative elements in the workbooks, see figure 9 and also as part of a Classroom Discussion Forum (CDF) (Hall, Woods, & Dautenhahn, 2004) session about MIXER held after the child had interacted with MIXER and completed workbook 2. The CDF session encouraged reflection and learning reinforcement; and qualitative data collection on the children's experience with MIXER. This activity included typical, in-role classroom activities, with a Q&A session, table discussions (small groups based on typical classroom seating plan) and general discussion about MIXER and the experience.

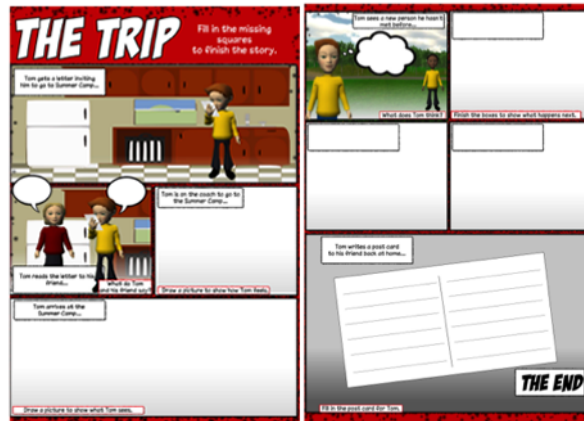


Figure 9: Qualitative Data Collection in the MIXER Workbooks

Questionnaire transformation is guided by user expectations, for example using images, colour, layout, interaction modality and style to transform instruments. For instance, if the evaluand is a space-based game, the user-role as space cadet, etc. then questionnaires can be given a space age look and feel, incorporated into the experience as part of the information needed to play the game. If our evaluand, was a child-focused tourist app providing facts and information about a stately home, we could ask the child to answer a quiz about their experience (e.g. showing retained learning), automatically receiving a badge on completion, thus resonating more clearly with user role as tourist.

Focusing on the user having an excellent experience both with MIXER and the evaluation, we found children appeared to enjoy responding to questions using a rating scale Rubie-Davies & Hattie, (2012) Inspired by child-focused hard-copy media aimed at recreational activity, such as comics, annuals and summer specials, we identified that children enjoyed: quizzes where they ‘discover’ something about themselves; activities with interactive elements, such as colouring-in, using stickers and limited text entry (e.g. completing empty speech bubbles); and questions incorporating puzzles, such as wordsearches, mazes, spot-the-difference, etc. With comic and activity books children expected a range of short, typically unrelated, complimentary, engaging and fun activities.

Although the media we sampled presented questions and activities with very different aesthetics, most comic books have the same elements, interspersed with, and themed by, the selling point of the comic, whether that is articles for pre-teen girls or more intrepid adventures for fans of Dr. Who. The techniques used to engage the user in comic books are relatively simple. Many activities incorporate vaguely relevant, but attractive, archetypal images (e.g. flowers and hearts; Dr. Who’s sonic screwdriver); others use colour blocking to link facts or present a group of related questions; motivators are also included such as directional arrows to move through an activity.

In the development of the props, we held questionnaire design workshops with children, both considering instrument design and to investigate whether providing the evaluation instruments in a comic book format was perceived as appropriate and engaging. However, the user role of school-based learner meant that the term comic book seemed inappropriate, conflicting and confused the role of learner with that of the role of comic book reader and fun-haver. Children

ENGAGING CHILDREN IN INTERACTIVE APPLICATION EVALUATION

instead suggested to us that we should call them workbooks, so that it was obvious that they were doing schoolwork. Using the term workbook also met with parent and teacher expectations, with many schools already using workbooks in the classroom.

Three workbooks were created for the pre-test (workbook 1), evaluation of immediate learning and experience of using MIXER (workbook 2) and the post-test. In many pre- post- tests, identical instruments (in content and format) are used. Instead, we wanted users to continue to engage with the questions rather than to feel a sense of déjà vu of having done all this before in Workbook 1. Thus, Workbook 3 presented a different appearance to incorporate the same questions and instruments, providing children with an engaging experience. Tables 1 and 2 provide the content of the workbooks with some sample pages in figures 10, 11 and 12.

Measures	Workbook 1 (Pre)	Workbook 3 (Post)
CQS	Which woodland animal are you? Designed as a quiz, with children rating which statements are like them and which not. Children are then identified as being a Badger, Fox or Deer, with all of the possible outcomes are constructively phrased and desirable for the children.	New People, New Places Children are given a series of images of mobile phones and asked to text Tom a number, 1 to 5, to tell him what they would do when making new friends
MESSY	New Friends The 20-item MESSY is designed to look like a puzzle, with children asked to help guide Ben to Barney. The cartoon bees are linked along a dotted line, interspersed with questions. The children move along this line 'helping' to get Ben back to Barney.	The MESSY was divided into three separate sets of questions: The Epic Quiz - children identifying on a scale how much things were like them. Friends: a series of questions providing learning about yourself Maze Days: children make their way through a maze answering the questions as they go.
Empathy Index	Yes or NO Presented as a comic strip, with each frame offer yes / no responses and the children following the arrow to the next box. Although this was the same box, whether yes or no was selected, no child has ever mentioned this.	Think Fast Think fast is a sticker activity where children are provided with YES and NO stickers to use to answer the questions.

Table 1: Pre & Post Tests in Workbooks 1 and 3

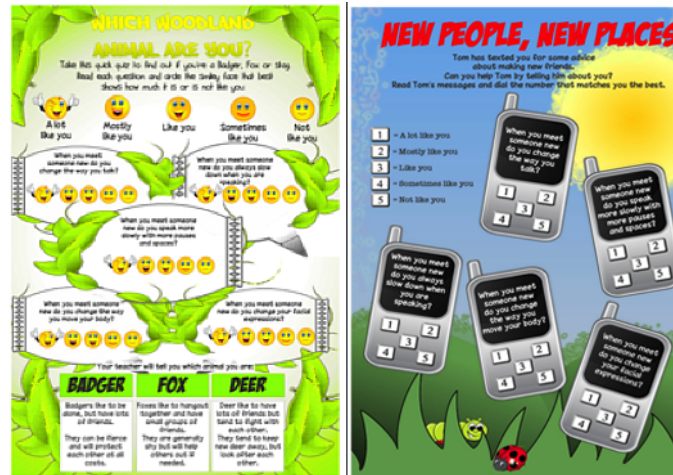


Figure 10: Worksheets 1 & 3 - MESSY



Figure 11: Workbook 1 & 3 - Bryant's Empathy Index

In the pre-test (Workbook 1), children were provided with some additional activities, including a maze to help Tom get to the summer camp (preparing the children for their interaction with Tom) and The Trip, see figure 9, a comic strip activity in which the children are given half of the story of Tom being invited to go to camp. Children are asked to complete the empty thought and speech bubbles and comic book squares. The children also write out a postcard for Tom to send home. The trip provides qualitative data on the children's perceptions of going to new places and meeting new people along with how they think another child may feel when away from home. A wordsearch was included as the final activity for workbook 1 and a colouring activity in workbook 3, so that any children who finished ahead of the other children would have something to do whilst the rest of the class finished.

Workbook 2, (see table 2), collects data related to children's immediate learning (near transfer); their narrative comprehension, empathic engagement; and their perspectives and views of the MIXER characters and experience. Workbook 2 addresses all four of the MIXER goals identified in the plot.

Instrument	Outline	Rating Approach
Who Wins?	Having used MIXER, children should have engaged with and have a deeper relationship with Tom than any of the other characters. It was expected that the majority of children would choose to put Tom in first place. This relates to the emotional and behavioural learning objectives.	Children place stickers of their 3 favourite characters onto a picture of a winner's podium.
Roving Reporter	Comprehension/opinion exercise to assess children's narrative comprehension and engagement with Tom. Higher scores for narrative show that children listened and paid attention to the story line. Positive responses equating to cognitive comprehension and deeper engagement with Tom.	Varied ratings from yes/no responses, and circling correct answer
True or False?	Features 8 questions. 6 questions address engagement and comprehension, i.e. they have a correct true / false answer, equating to emotional, behavioural and cognitive learning. 2 questions gather children's opinions of the rule conflict reflecting the cognitive and behavioural learning outcomes.	The children use 'True' or 'false' stickers to answer the questions.
MIXER views	Features questions on user experience with MIXER (e.g. appropriateness of duration, desire to use MIXER again, etc.), equating to experiential learning.	Children circle one of the given responses.
What do you think?	Evaluates usability (e.g. voices, text, etc.) and experience (e.g. who explained the rules the best) of the MIXER application, relating to experiential learning.	Selections and Yes/No responses
iPad Page	Provides an evaluation of the interaction approach. e.g. 'Do you think the game on the iPad was easy to use/not easy to use exciting/dull'	5-point Likert scale represented as faces.

Table 2: Engagement Experience Questionnaire



Figure 12: Workbook 2 - Engagement with Tom; immediate Learning Assessment; Interaction Modality Evaluation

Our approach to data collection transformation has had a significant impact on the user's perception of what they are doing. Users are usually unaware that they are completing questionnaires, being assessed on their learning or participating in a focus group for example, as the props that they are engaging with are

embedded and just part of their in-role experience (Hall et al., 2013). We recognize that to any experienced evaluator they are clearly questionnaires and focus groups, however, this is not the user perception, with the instruments masked through adhering to user role and meeting the user expectations of that role. This was achieved with MIXER, with all props reinforcing the user role of school-based learner and ensuring that the children were having an excellent experience in that role. Children eagerly engaged with the workbooks, with 100% self-regulated completion. Children were very positive about all elements of the MIXER Summative Evaluation, with some children saying that they enjoyed the workbooks more than interacting with MIXER. From observation and discussion, throughout the MIXER Summative Evaluation children appeared to be as engaged with the evaluation battery as with the evaluand.

4.1.3 Script - Protocol & Procedure

The Transmedia script provides the experimental protocol and procedure for the evaluation event. The script implements the plot, ensuring that each plot node can be achieved, whether that be to engage in training, interaction or evaluation, whilst the user role (in-role expectations and user experience expectations) can be maintained and R&D requirements met through appropriate props and activities.

The script unites the various elements of the evaluation into a single coherent narrative. The script of the event incorporates all of the user experience, including the initiation of an event, recruitment for the event and the completion of the event, typically a final engagement with the evaluators or the information relating to the next event. The finish point of a Transmedia Evaluation reinforces the user role and the expectation that their experience has been of value to the researchers.

Depending on user role, the script may have a theatrical focus, placing evaluators and researchers into in-context roles, for example as non-player characters in game evaluations with specific utterances. Or it may leave evaluators in a primarily researcher role, to cope with software failings for example. The script typically requires the evaluation team to explain certain issues or to say specific texts (particularly if the evaluation team take in-evaluation experience roles) and assumes a positive, constructive and upbeat approach from the evaluators. This upbeat approach is a vital part of evaluation, especially for children. The assumption and basis of the script in Transmedia Evaluation is that whatever role the evaluator is in, they will improve the experience for the user.

The MIXER Summative Evaluation script placed users firmly in their role as school based learners having a great time using educational technology. Initial recruitment of children involves explaining the evaluand, evaluation battery and researcher role through a script that highlights that in big technology enhanced learning projects we need to get users (e.g. them) to try out the learning materials. More detail explains that we are University researchers working on Personal, Social and Emotional (PSE), and we want the children to try out our technology and see if it works. Telling the children that the experience will include using an iPad gives it considerable appeal for the target age group. In the recruitment phase, the script clearly identified that the children were engaging

in an evaluation, with the ethics forms and the information accompanying them clearly stating that the purpose of the experience was a user evaluation.

Workbook 1 was completed during the pre-test. At the end of the session, the researchers explained that children would interact with some new learning technology using an iPad in their next session. In the script, is the instruction for the evaluators to 'prime' users to expect a good experience, to look forward to their next encounter, and to excite them about what will happen next.

Children interact with MIXER during the test phase followed by completion of, workbook 2. They also engage in a learning reinforcement session and qualitative data collection related to children's immediate learning, their engagement with MIXER and their satisfaction (enjoyment) with the interaction. The event concludes with the evaluators explaining the next meeting and priming children's anticipation.

Although the script initially incorporated a finish point where we returned to the school and provided results, the school requirements (related to Christmas Plays and seasonal events) meant that we could only realistically have 3 sessions, requiring the post-test to also provide the completion point.

4.1.4 Rehearsal & Refinement

Transmedia Evaluation requires an iterative, incremental method, with all elements of the user experience, such as the role, instruments, approaches and activities developed with design input from users and then piloted with representative users. With the focus of the evaluation being the provision of data to the R&D team we also pilot the data capture and analysis approaches, aiming to ensure that R&D expectations and requirements are met. Rehearsal is used to develop the evaluation instruments and experience in parallel to the development of the evaluand. As the evaluand develops from low-fi (pen and paper) versions to hi-tech (implemented system) so to does the evaluation.

The user is required to suspend disbelief, interpreting and achieving all aspects of their experience whilst immersed and engaged in their role. Rehearsal identifies points where immersion may fracture identifying aspects of the experience that need improvement. All this extensive piloting identifies problems that can be resolved or reduced through appropriate experience design. Although rehearsal happens throughout the design of the experience and may frequently be targeted at specific elements, such as the instrument to capture the data, it is critical to regularly have rehearsals of the entire experience to ensure that the event works as a whole performance and not just in parts.

The MIXER Summative Evaluation was the culmination of 3 years of work for the R&D team. The 3 workbooks and the CDF had been extensively piloted with users, with an initial workbook design piloted in the first year of the project. The Summative Evaluation had a large-scale pilot as the final rehearsal with results highlighting a significant flaw in our experience design through placing the discursive and qualitative activities at the end of the experience, rather than at the end of the interaction (Aylett et al., 2014). R&D input and discussions with teachers highlighted the need to change the experience design to reinforce intercultural learning soon after the MIXER interaction, rather than after the entire experience. Children's response to MIXER was the expectation that they

would get a chance to talk to each other and us about MIXER straight away rather than a week later. This also met with R&D team expectations, as immediate user response to the interaction was more valuable than their memories of the experience.

4.1.5 Performance (Transmedia Evaluation Event)

A Transmedia Evaluation event incorporates all of the user experience, from recruitment to completion. Singular one-off experiences (e.g. interaction followed by user satisfaction questionnaire or learning assessment) or longitudinal designs are supported (e.g. as with multiple episodes as required by a pre- post-test design).

All of the elements (plot, role, props and script) feed into the event phase during which the participants and evaluators are in role and the evaluation occurs. The event is the shortest phase of an evaluation, with the procedure, instruments, data capture and evaluation all prepared, rehearsed and refined. After the event, the data is prepared for analysis following the specified protocols and analysis begins.

The script provides both the protocol and instructions to the evaluators for how the evaluation is to occur, providing the detail underpinning the plot. With MIXER, see figure 8, the script has the following nodes:

- **Start:** With the MIXER summative evaluation we met with children prior to the Pre-test, for a brief 10-minute session at the beginning of the school day. Our instructions were to introduce ourselves, the project and the experience, with the bottom line being to enthuse the users about their experience. We briefly explained that they would be completing some workbooks, we showed them these from the front of the class, and that they would get to use MIXER where they would meet Tom, some images shown. We told them we'd be coming three times (pre, interaction, post). The ethics documentation and experience information was provided to children and the school provided us with the completed ethics forms prior to the pre-test.
- **Pre:** The pre-test involved the children being given the workbook and asked to complete it. Children worked individually on their group tables. At the end of the pre-test the children were briefly told about what the next session would include. 100% self-regulated completion was achieved. The final activity of workbook one was a time filler, a word search used for those children who completed the questionnaire quickly.
- **Test:** the limited equipment and school requirements resulted in children interacting with MIXER (individually) in small groups of 4 in the library near to the classroom. Children not using MIXER were engaged in class-based activities with the teacher. Once all children had interacted with MIXER and completed workbook 2, the learning reinforcement session stimulated debate and discussion, naturally moving through the various qualitative questions, relating to the children's experience of MIXER
- **Post-Test:** At the beginning of the school day the workbooks were distributed and completed. After completion, the evaluators then explained what would happen with the results, highlighting the value of

the children's input for understanding technology enhanced learning, hoped that they had had a good time and thanked the children, providing each child with an eCute mascot.

4.1.6 Outputs & Review

The review phase assesses the event in relation to the outputs, that is the results achieved and their use by the R&D team. The review phase of the Transmedia Evaluation also provides the evaluation team with the opportunity to reflect upon the evaluation, considering what aspects of the event went well and what could be improved. This then feeds into the design of subsequent evaluations, identifying successful activities.

The MIXER summative evaluation was a very successful experience for all concerned. From an R&D perspective, the data was complete, of high quality and from engaged participants. An overview of the results is presented in table 3.

Learning Goal	Learning Objective	Near Transfer Learning Goal Achieved? (EEQ & CDF measurements)	Far Transfer Learning Goal Achieved? (CQS, Bryant's Empathy, Messy)
Emotion Goals	Be able to recognise emotions (for example fear and anxiety) when dealing with the strange behaviours of another group.	<p>YES - Children showed a preference for the characters that they interacted the most with, and those that displayed the most narrative.</p> <p>YES - Children wanted to be friends with 'Tom' and felt that he had listened to them, which demonstrates the ability of children to recognise their different emotions during MIXER towards Tom (in-group) compared to their emotions towards the 'yellow' team (out-group).</p> <p>YES - Children demonstrated high levels of engagement with the MIXER software. They thought MIXER was fun, and many children wanted the interaction to be longer.</p> <p>YES - Nearly 90% of children expressed a desire to use the MIXER software again.</p>	<p>NO - Children's empathy levels remained constant between pre-post-test after interacting with MIXER.</p> <p>NO - Messy (social interaction ability) scores were unchanged between pre-post-test after the MIXER interaction.</p>
Cognitive Goal	Start learning the specific practices and values of that group.	<p>Some evidence - Ability to comprehend the 'rule-change' between the red and yellow teams demonstrated by some children. But, nearly 50% of children did not appreciate the 'cultural differences' of the 'yellow' team rules just being different and not cheating.</p> <p>YES - High levels of comprehension as children understood the events and progression through the game of Werewolves.</p>	<p>YES - Children's CQS scores were higher at post-test after the MIXER interaction. Provides some evidence that children had started to learn conceptually about the values and attitudes of the MIXER characters.</p>

		YES - Children understood the MIXER application and its narrative.	
Behaviour Goal	Being fully present in attending to others verbal and non-verbal messages.	<p>YES - Children who wanted to be friends with Tom also felt that they had helped Tom and that Tom had listened to them.</p> <p>YES - Children who believed they had helped Tom believed Tom knew what he was doing, felt that Tom was good at Werewolves.</p> <p>YES - Positive association between children believing Tom had listened to them and Tom having fun in Werewolves.</p>	YES – children’s CQS and MESSY scores were higher after interacting with the MIXER software at post-test (T2). This suggests children started to adapt and modify their behaviour/facial expressions/vocalisations to the novel MIXER scenarios.

Table 3: Summary of the main results from the MIXER Transmedia Evaluation

Qualitative data collected in the learning reinforcement classroom session highlighted children had really engaged in the experience provided to them via the Transmedia Evaluation. Children were enthusiastic about all elements of the MIXER Summative Evaluation, including the interaction, the workbooks and the class-based discussion. All the instruments incorporated into the workbooks were 100% completed. At no stage during the completion of the workbooks did any child ask for help or support in completing the activities. Not only did the evaluation identify that children enjoyed using MIXER, but additionally that children successfully engaged in experiential learning, empathically engaged with the characters and enjoyed their experience of the evaluation.

The MIXER Summative Evaluation resulted in publishable outcomes. Results enabled the R&D team to highlight that children exhibited both near and far transfer, meeting the learning goals of the eCute Intercultural Learning Framework and contributing to the Excellent rating achieved by the eCute project in its final review with the European Commission.

5 Discussion

Transmedia Evaluation provides a different approach to evaluation than that commonly seen in the design of evaluation experiences for children. Instead of evaluation being a discrete task performed with traditional approaches, Transmedia Evaluation creates an engaging, coherent, integrated experience. All elements of the event are visible to the user as part of a consistent experience, facilitating the user in adopting and maintaining their assigned role in the event, see figure 13. Transmedia Evaluation aims to provide a methodology that represents and respects children’s interests.

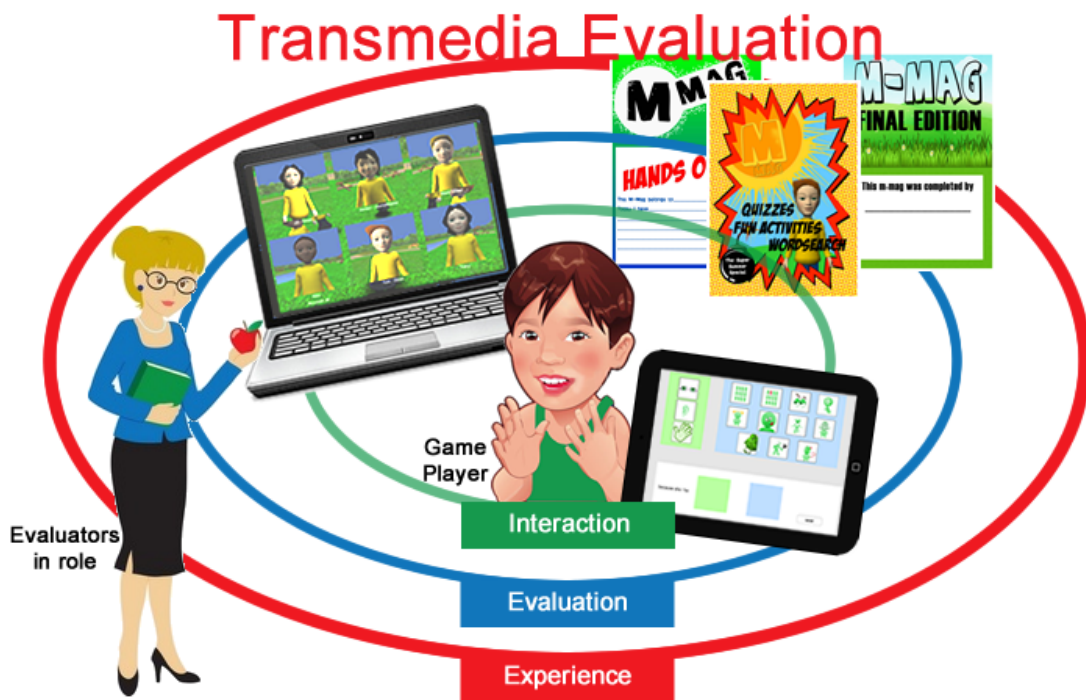


Figure 13: Child's Experience of the MIXER Summative Evaluation

Transmedia Evaluation enables the generation of appropriate results for the R&D teams and research requirements to be met. The plot of a Transmedia Evaluation Event provides the purpose of the evaluation, defining the constraints and requirements that ensure that the appropriate data is gathered. In the MIXER Summative Evaluation detailed in this paper, the plot nodes were recruitment, learner baseline measurement, interaction, evaluation of learner's immediate learning and engagement with MIXER, reinforcement and reflection of intercultural learning and post-test measurement. Specific instruments included in the pre- and post- tests were Bryant's Empathy questionnaire, the MESSY measurement of social interaction skills and Ang's Cultural Intelligence Questionnaire, whilst immediate learning was assessed by the CEQ and CDF. Ethics requirements were met (requiring children and parents to complete relevant forms) and included the collection of quantitative (using identified measures), qualitative (using open questions in written format and classroom based discussions), observation (recordings of children engaging with MIXER and the evaluation) and interaction (children's interactions with Tom) data. This sounds like an arduous and disengaging plot, however, by placing the plot within a script where the user played an engaged, motivated school based learner having an excellent experience, completely renegotiates the evaluation space. Perceiving an evaluation as adding to the evaluation experience, rather than simply assessing completely changes the dynamic of the evaluation process.

The plot provides the structure and the key elements of the experience. It is all too easy to lazily respond to research questions by providing a traditional, disruptive evaluation experience that meets R&D requirements but shows little thought for either the user or the evaluand. However, by responding to everyday user expectations (which for children are ALWAYS to have an excellent

experience) and the expectations that the user role creates (adding context and/or value to the evaluand), as outlined for the MIXER Summative Evaluation, a vastly different experience can be achieved for the user. Rather than the child experiencing the role of subject completing arduous evaluation instruments, instead they are experiencing an excellent lesson, designed and crafted around their engagement.

Applying the Transmedia Evaluation methodology to large scale R&D evaluations with children naturally highlights the need for multiple methods to collect data. This not only meets the R&D requirements for a variety of data types, including logged, quantitative and qualitative data. But, further it meets children's expectation of a variety of interesting, diverse activities rather than responding to a group of semi-identical questionnaires for 20 minutes. In the MIXER Summative Evaluation mixed methods included logged data of the child's interactions with Tom providing responses to Theory of Mind questions; mainly quantitative but also written/ drawn qualitative data in the workbooks; and the CDF verbally assessing immediate learning and the children's experience.

The dominance of questionnaire measures and their blanket acceptance within the research community essentially demands that large R&D evaluations of interactive systems incorporate questionnaire instruments. Repetition in questions and assessment style (in structure / approach rather than content) seems contrary to our vision of evaluation. Yet, as we have detailed in the MIXER Summative Evaluation, initial versions of the questionnaires did present with a similar style, look and feel. Our transformation of these instruments was effective as none of the children who participated in the evaluation were even slightly aware that they were completing questionnaires. Nor could children be prompted to discuss the experience as anything but engaging with a fun workbook. All of the workbooks (357) were 100% complete, with variation in answers, identifying that children were engaged in providing optimal answers, rather than satisficing or adhering to acquiescence bias. Many of the children would have enjoyed further workbooks and mentioned this in the CDF. This is a significant outcome for us as we are unaware of any other evaluations where child users have asked to fill in more questionnaires.

Transforming both qualitative and quantitative data capture instruments into engaging elements of the user experience is achievable. It requires evaluator investment in understanding user expectations by reviewing literature and experiences and from exploring the appropriateness of the evaluation with the users themselves. Including children in the evaluation design is essential, instruments and techniques must be piloted and children's ideas incorporated ensuring that the expectation of having a great experience is realized. Children's input can be insightful and improve our approaches. For example, with the MIXER Summative Evaluation children's input resulted in the instruments being called workbooks and for the focus group to be reconceptualised as a Q&A session with raised hands as the best way to gain class-based feedback.

In Transmedia Evaluation we strongly advocate the use of established measures. It is the transformation process that is central to providing an engaging evaluation experience, typically reducing and refining them and almost always changing how they are presented to the user. Clearly, our transformation of

measures may have an impact on the validity and reliability of the chosen instrument. However, without doubt using existing measures and techniques will increase the quality and reliability of the data, rather than a researcher creating a new (usually barely piloted) instrument. In practice, the transformation of instruments is relatively straightforward adhering to the more achievable aspects of engagement such as age appropriateness, length of experience, and aesthetic appeal. Graphical approaches, such as attractive layout and relevant images, significantly change how questionnaires are received. Incorporating 'story' elements from the evaluand into the design and simply improving the appearance to include age-appropriate design improves the evaluation experience for children.

However, with Transmedia Evaluation, the central intention is to engage the user throughout the evaluation experience with equal immersion through all process steps, not just to provide age appropriate instruments nor prettify the evaluation battery. Although well-designed instruments do increase engagement, to ensure immersion and to prevent an experience rupture or dislocation, more is needed than cosmetic improvement. It is essential that the experience is considered at a meta-level, creating an overarching theme, consistent with the child's expectations and the narrative of the evaluand. Only then, do the instruments become analogous to transmedia, encircling the interaction and building the story and experience for the user.

Placing the child at the centre of the evaluation in a clearly defined user role provides inspires the design of props and scripts. With user role as the guiding principle, instead of designing data collection tools we are designing elements of an experience. The props fit with the role and evaluand, reinforcing the sense of a single, coherent experience for the child. The script seamlessly integrates all elements of the user's experience, including recruitment, training, interaction and evaluation, binding together the various elements into a single coherent narrative. For children, Transmedia Evaluation creates an effective, enjoyable evaluation, where interaction and data collection support the child's immersion into a coherent, engaging and enjoyable experience.

Transmedia Evaluation provides an optimal experience both for children and R&D aims with regards to the quality and richness of data collected. As highlighted above, a potential shortcoming of Transmedia Evaluation may be that issues surrounding instrument validity and reliability are not fully adhered to. Future research should aim to carry out comparative Transmedia Evaluation Studies with traditional pencil-paper approaches to determine whether this is the case. This paper has clearly presented that Transmedia Evaluation provides an engaging and immersive experience for children. Future studies should extend the Transmedia Evaluation approach to include other participant cohorts with adults and other research disciplines.

6 Conclusions

This paper clearly demonstrates that by placing children at the centre of evaluation design and meeting their expectation of enjoyment, we can both respect and represent their interests in the evaluation process. Applying the Transmedia Evaluation process to the MIXER Summative Evaluation generated relevant, high quality results for the R&D team, through transforming and

integrating the evaluation into an excellent, coherent experience for the child. Transmedia Evaluation provides an innovative, effective evaluation methodology that enriches the evaluation process, generating high quality data whilst providing a different, enhanced experience for children.

References

- Ang, S., Van Dyne, L., Koh, C., Ng, K. Y., Templer, K. J., Tay, C., & Chandrasekar, N. A. (2007). Cultural Intelligence: Its Measurement and Effects on Cultural Judgment and Decision Making, Cultural Adaptation and Task Performance. *Management and Organization Review*, 3(3), 335–371. doi:10.1111/j.1740-8784.2007.00082.x
- Aylett, R., Lim, M. Y., Hall, L., Endrass, B., Tazzyman, S., Ritter, C., ... Kappas, A. (2014). Werewolves, Cheats and Cultural Sensitivity. In *2014 International Conference on Autonomous Agents and Multiagent Systems* (pp. 1085–1092). Paris, France: International Foundation for Autonomous Agents and Multiagent Systems.
- Babbitt, B. (1989). Questionnaire construction manual annex. Questionnaires: Literature survey and bibliography. *Operations Research Associates*, (June).
- Bryant, B. (1982). An index of empathy for children and adolescents. *Child Development*, 53(2), 413–425.
- Buckleitner, W. (1999). The State of Children’s Software Evaluation—Yesterday, Today, and in the 21st Century. *Information Technology in Childhood Education Annual*, 1999(1), 211–220.
- Endrass, B., Hall, L., Hume, C., Tazzyman, S., & Andre, E. (2014). A Pictorial Interaction Language for Children to Communicate with Cultural Virtual Characters. In *16th International Conference on Human Interaction* (p. in press). Heraklion, Greece.
- Endrass, B., Hall, L., Hume, C., Tazzyman, S., Andre, E., & Aylett, R. (2014). Engaging with virtual characters using a pictorial interaction language. In *CHI’14 Extended Abstracts on Human Factors in Computing Systems* (pp. 531–534).
- Gomez, J., & Pulman, S. (2012). In Ridley Scott’s “Prometheus,” the Advertising Is Part of the Picture. Retrieved December 12, 2012, from <http://adage.com/article/digitalnext/ridley-scott-s-prometheus-advertising-part-picture/233452/>
- Hall, L., & Hume, C. (2011). Why Numbers, Invites and Visits are not Enough: Evaluating the User Experience in Social Eco-Systems. In *SOTICS 2011, The First International Conference on Social Eco-Informatics* (pp. 8–13).
- Hall, L., Jones, S., Aylett, R., Hall, M., Tazzyman, S., Paiva, A., & Humphries, L. (2013). Serious Game Evaluation as a Metagame. *Journal of Interactive Technology and Smart Education*.
- Hall, L., Woods, S., & Aylett, R. (2006). FEARNOT! Involving Children In The Design Of A Virtual Learning Environment. *Journal of Artificial Intelligence and Education: Special Issue on Learner Centred Methods for Designing Intelligent Learning Environments*, 16(4), 237–251.
- Hall, L., Woods, S. and Dautenhahn, K. (2004) FearNot! Designing in the Classroom, in Fincher, S., Markopoulos, P., Moore, D., & Ruddle, R. (Eds.) *People and computers XVIII: Design for life*, Vol. 2 (Proceedings of HCI 2004), Springer.

ENGAGING CHILDREN IN INTERACTIVE APPLICATION EVALUATION

Hall, L., Woods, S., & Hall, M. (2009). Lessons Learned Using Theory of Mind Methods to Investigate User Social Awareness in Virtual Role-Play. *Journal of Human Technology, Special Issue on The End of Cognition*, 5(May), 68–89.

Jenkins, H. (2011). Transmedia 202: Further Reflections. Retrieved December 10, 2012, from http://henryjenkins.org/2011/08/defining_transmedia_further_re.html

Jensen, J. J., & Skov, M. B. (2005). A review of research methods in children's technology design. In *Proceeding of the 2005 conference on Interaction design and children - IDC '05* (pp. 80–87). doi:10.1145/1109540.1109551

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996, 29–44. doi:10.1002/ev.1033

Matson, J. L., Neal, D., Fodstad, J. C., Hess, J. a, Mahan, S., & Rivet, T. T. (2010). Reliability and validity of the Matson Evaluation of Social Skills with Youngsters. *Behavior Modification*, 34(6), 539–58. doi:10.1177/0145445510384844

O'Brien, H. ., & MacLean, K. . (2009). Measuring the User Engagement Process. *Advances*, 1–6. Retrieved from http://faculty.arts.ubc.ca/hobrien/files/OBrien_MacLean_2009_Measuring_the_User_Engagement_Process.pdf

O'Brien, H. ., & Toms, E. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1), 50–69. doi:10.1002/asi.21229.1

Ólafsson, K., Livingstone, S., & Haddon, L. (2013). Children's use of online technologies in Europe: a review of the European evidence base.

Parker, J., & McDonald, R. (2014). Stories are More than Paper: Using Transmedia with Young Adults.

Read, J. C., & MacFarlane, S. (2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. *Proceeding of the 2006 Conference on Interaction Design and Children IDC 06*, 81. doi:10.1145/1139073.1139096

Read, J. C., Markopoulos, P., Par, Eacute, S, N., Hourcade, J. P., & Antle, A. N. (2008). Child computer interaction. *Proceedings of ACM CHI 2008 Conference on Human Factors in Computing Systems*, 2, 2419–2422. doi:10.1145/1358628.1358697

Rubie-Davies, C. M., & Hattie, J. A. C. (2012). The dangers of extreme positive responses in Likert scales administered to young children. *The International Journal of Educational and Psychological Assessment*, 11, 75–89.

Sapouna, M., Wolke, D., Vannini, N., Watson, S., Woods, S., Schneider, W., ... Aylett, R. (2010). Virtual learning intervention to reduce bullying victimization in primary school: a controlled trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 51(1), 104–12.

Swiderska, A., Krumhuber, E., Kappas, A., Degens, N., & Hofstede, G. J. (2011). *D2.1 Preliminary cultural learning interdisciplinary framework*. Retrieved from <http://ecute.eu/wp-content/uploads/downloads/2012/08/D2.1Preliminary-cultural-learning-interdisciplinary-framework.pdf>

Weiler, L. (2012). Building Storyworlds. Retrieved December 11, 2012, from www.lanceweiler.com

Zaman, B., Vanden Abeele, V., Markopoulos, P., & Marshall, P. (2012). Editorial: The evolving field of tangible interaction for children: The challenge of empirical validation. *Personal and Ubiquitous Computing*. doi:10.1007/s00779-011-0409-x

List of Tables

Table 1: Pre & Post Tests in Workbooks 1 and 3	15
Table 2: Engagement Experience Questionnaire	17
Table 3: Summary of the main results from the MIXER Transmedia Evaluation	22

List of Figures

Figure 1: Standard 'disruptive' evaluation approach.....	3
Figure 2: Alex and Lisa, characters from MIXER	5
Figure 3: Scenes from MIXER	5
Figure 4: Fragment of Pictorial Interaction Language.....	5
Figure 5: A Transmedia Evaluation Event	7
Figure 6: Basic, Better and Best materials from early stage evaluation	8
Figure 7: Transmedia Evaluation Framework	9
Figure 8: MIXER Plot for Summative Evaluation	11
Figure 9: Qualitative Data Collection in the MIXER Workbooks	14
Figure 10: Workbooks 1 & 3 - MESSY	16
Figure 11: Workbook 1 & 3 - Bryant's Empathy Index	16
Figure 12: Workbook 2 - Engagement with Tom; immediate Learning Assessment; Interaction Modality Evaluation.....	17
Figure 13: Child's Experience of the MIXER Summative Evaluation.....	23

A Pictorial Interaction Language for Children to Communicate with Cultural Virtual Characters

Birgit Endrass¹, Lynne Hall², Colette Hume², Sarah Tazzyman²,
and Elisabeth André¹

¹ Human Centered Multimedia, Augsburg University,
D-86159 Augsburg, Germany
{endrass, andre}@hcm-lab.de

² University of Sunderland,
Sunderland, SR6 0DD, UK
{lynne.hall, colette.hume, sarah.tazzyman}@sunderland.ac.uk

Abstract. In this paper, we outline the creation of an engaging and intuitive pictorial language as an interaction modality to be used by school children aged 9 to 11 years to interact with virtual characters in a cultural learning environment. Interaction takes place on a touch screen tablet computer linked to a desktop computer on which the characters are displayed. To investigate the benefit of such an interaction style, we conducted an evaluation study to compare the pictorial interaction language with a menu-driven version for the same system. Results indicate that children found the pictorial interaction language more fun and more exciting than the menus, with users expressing a desire to interact for longer using the pictorial interaction language. Thus, we think the pictorial interaction language can help support the children's experiential learning, allowing them to concentrate on the content of the cultural learning scenario.

Keywords: Interaction Design, Interaction Modality, Virtual Agents, Culture, User Experience.

1 Introduction

While traditional learning systems provide conventional interaction devices such as mouse and keyboard, especially for child users intuitive interaction is important to provide an engaging experience. Menus provide bound and restricted interaction choices, possibly limiting the user's perceived freedom in their interactions. Free text input can be desirable, however, due to technical constraints such as limited support of vocabulary and grammar this is hard to realise. Further, children's keyboard skills are often not fully developed compared to adults, reducing children's abilities to express themselves. Recent paradigm shifts towards more natural user interfaces, based on either direct touch or three-dimensional spatial interaction [1] provide interesting alternatives to increasing user engagement, particularly for children. One of the most often stated benefits is the

view that interacting with an application through directly touching graphical elements is a more "natural" or "compelling" approach than working indirectly with a mouse or other pointing devices [2].

In this paper, we investigate a game play interaction modality designed for, and with children. We present a pictorial interaction language using touch-based gestures on a tablet computer that allows children to interact freely with characters displayed on a different screen. The interaction is developed for use in playing a serious game in which children communicate with a virtual character to learn about resolving a cultural conflict.

2 Background

This paper focuses on the development and evaluation of a pictorial interaction language for children aged 9 to 11 years. This interaction modality is part of the eCute project [3], which aims to create and encourage cultural awareness among children.

In the MIXER showcase [4], the user plays the role of an invisible friend to provide advice and support to a virtual character, called Tom. The narrative of the MIXER application centres on Tom visiting a summer camp where he meets a group of characters that he knew before. With this group, Tom plays a game called Werewolves (see [5] for a description of the rule set). In the game, each player is assigned a role, as either a werewolf or a villager. The aim of the game is to deduce which character in the group is the werewolf, before the werewolf kills all of the villagers. Several times during the game, Tom asks for the user's advice. After playing for a while, Tom leaves the first group of children and meets a different group that he has not met before. In this group, Tom and the user are confronted with crucial changes to the rule set by which the game is played; this leads to a critical incident and a potential conflict situation.

To create a novel, engaging and effective learning experience we aim to develop an interaction modality that is both intuitive and engaging for children of the target age group. A pictorial interaction language was identified as a solution to the problem of creating a novel and universal interaction experience.

3 Related Work

We think that finding novel and intuitive interaction modalities for educational systems is a crucial task. Sali and colleagues [6] investigated three different dialogue approaches for game interfaces. They found that users prefer a natural language interface over interfaces that allow users to select sentences and interfaces that make use of an abstract response menu interface. However, some users had problems with the natural language interface because they found it hard to figure out what to say in a particular situation. Compared to adults, this problem may be magnified for children. We encountered similar issues in former work [7] where children interacted with a virtual learning environment using typed text input. The interaction choice for natural language input resulted in

several problems including recognition problems for the software coupled with the difficulty and time required for children to express themselves in typed text. We think that by using a pictorial interaction language the disadvantages of text input are reduced, whilst retaining a large degree of interaction freedom.

Pictorial languages are commonly used with children in the field of augmentative and alternative communication, e.g. in communication training for autistic children [8], [9]. Widget symbols (e.g. [10]) also find their usage on websites that provide understanding and communication for people who find reading text difficult, e.g. [11]. Their potential for intuitive communication is gaining ground for non-disabled children as well. For example, a pictorial language is used on CBBC (children's television channel) in the UK to facilitate communication. We thus see great potential in using a pictorial language as an intuitive interaction modality to communicate with virtual characters in learning environments as well.

Researchers have found that in the field of human computer interaction using a visual style of expressing oneself helps to motivate children to complete creative and challenging tasks, such as telling stories [12], or learn computer programming using storytelling environments [13].

To overcome the language barrier in inter-cultural communication, Takasaki and Mori [14] describe a communicator that was developed for children of different cultural backgrounds to be able to talk to each other using pictogram communication. This was reported to be an effective and practical user interface design method with children. In a similar manner, we aim to design a pictorial language for children to enable communication with a virtual character.

4 Design and Realisation

With the intention of improving both engagement and user experience for 9-11 year olds, we use an Apple iPad as the interaction device in combination with a pictorial language as interaction modality, provided as an extension to a desktop-based system connected via Wi-Fi.

4.1 System Setup

Figure 1 shows an overview of the setup including a child using it. The user can observe what happens in the virtual environment on the screen of the desktop while interaction takes place on the tablet computer. In this way, information that should only be visible to the user is shown on the tablet computer, while the environment with the virtual characters is visible for everyone. This supports the impression of being an invisible friend whose actions cannot be seen by the other characters involved in the gameplay.

As the focus of the present study was to test the suitability of the pictorial interaction language, it was tested with an early version of the MIXER game holding a virtual friend character that is involved in a fictive game with a group of characters (running in the AAA application [15]). During the game, the friend

character asks the user for advice several times. In each case, the character leaves the group, comes closer to the screen and updates the user on what happened in the game. Depending on the context of the question, different icons are provided on the tablet computer to construct the answer message in a pre-structured "grammar", by e.g. combining an action and an emotion. After hitting the send-button, the friend character reacts to the message and returns back to the group of other characters.

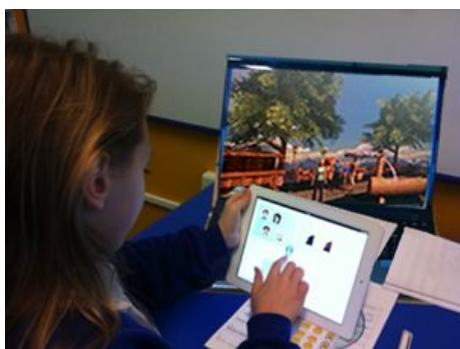


Fig. 1. Example setup, with a child using the pictorial interaction interface on an iPad

4.2 Interaction Modes

We designed two different advisory modes for interaction with the virtual friend character in the MIXER game:

- "During game advisory mode" to support the friend character during game-play;
- "Critical incident advisory mode" to deal with the critical incident after playing with a different group of characters that play the same game with different rules.

In this paper, only the advisory modes that occur during the game were investigated. Therefore, we identified four different advisory modes that describe standard situations for the Werewolf game: (1) Questioning who the werewolf is, (2) Reasoning why somebody is the werewolf, (3) Reacting if somebody else is being accused, (4) Reacting if oneself is being accused.

Depending on the context of the game, the advisory modes can either be used alone or combined to simulate a longer conversation between the user and the friend character. For example, after a character has been accused of being a werewolf (3) the friend character could ask who else could be a candidate (1) and why the user thinks so (2). Interaction is managed in a question and answer style, with the friend character asking for advice and the user answering by constructing a message.

4.3 Vocabulary Selection

We had to create a language that would fit our purpose of communicating with a virtual character that was playing a game of werewolves with other virtual characters. As this was a very specific requirement, we could not, for example, acquire a set of validated open source icons. It was necessary to create and test our own icon set. The first stage in the creation process was to investigate the language used while playing the werewolve game. Taking into account the rules of the game, some words were obvious, such as 'You', 'They', 'Accuse', 'Defend', 'Werewolf' etc. We recruited a total of 70 children (aged 9 to 11) who played the real world werewolf card game in small groups. The games were recorded and transcribed. In total, we identified 60 frequently used words, such as "I **accused** her because she **looks suspicious**" or "he's being too **quiet**". These words were later grouped, e.g. emotions or actions and structured in a way to match the identified interaction modes.

4.4 Icon Design

Besides following the design standards of ISO/IEC 11581 by using a consistent size, icon behaviour, and a similar design for all of our icons, a challenge was to design the icons to be sufficiently intuitive for children to construct meaningful messages.

In total, a set of over 60 icons was required for the pictorial interaction language based on the study mentioned above. The majority of the icons show a small green character. This character was shown in different positions to convey the different action states that were identified. For example, for the word 'calmly' the character was shown in a meditative pose. The colours green and red were used in the icons to convey positive and negative respectively. The icons were designed intuitively, by using what seemed to be the most appropriate visual representation for children of each word. However, what is obvious to a team of researchers is clearly not always going to be obvious to a child. Thus, we conducted a small study with 30 children to test their comprehension of the icons. We began by introducing and playing a short game of werewolves, which gave the children a context in which to discuss the meaning of the icons. The children were given activity sheets with pictures of the icons, and then asked to think about the game they had played and to try and work out what each of the icons meant. Following the game we held small focus group activities during which the children were shown the icons again and asked to discuss what the icons represented. This gave the research team useful qualitative information about children's views of the icons and their design. The icons that were not easily understood by the children became part of the next activity in which the children themselves helped to redesign the icons. These were used to develop the final icon set. The focus group activity was repeated with the final icon set with a further 25 children at a different school, during which all icons were successfully identified by the children. Figure 2(left) shows a small subset of the icons designed for our pictorial interaction language along with their intended meanings.



Fig. 2. Left: Example icons with intended meanings; Right: Interaction interface on the iPad

4.5 Interface Design

For the interface shown on the iPad, groups of icons are provided, while one icon of each group can be selected to form a sentence in a pre-structured grammar, e.g. by combining an action with an emotion. Figure 2 (right) shows the iPad with an interactive screen of the third advisory mode. Different coloured views contain the different groups of icons. Icons are moved by touching and dragging them across the screen. The white area on the lower part of the screen holds the message that the user constructs, providing empty views of the same colour of the group of icons that can be selected. The simple colour code helps the user understand that an icon of each provided group should be selected and moved to the corresponding area. Additionally, icons are automatically attached to the correct position (centre of same coloured area) as soon as they are moved into the user sentence area. In case an icon was selected for that area before it is replaced and the former is popping back to its initial position. Thus, only well formed sentences can be produced by the child, not allowing grammatically incorrect or nonsense sentences that would be uninterpretable for the system.

An example of a standard situation in the Werewolf game includes questioning why another player might be a werewolf (2). To answer this question, an action and a reason can be combined by the user. To help the user understand what kind of answers can be created, the message is initialised by the words "because he / she", followed by two different coloured views relating to actions and reasons respectively. Using the icons shown in Figure 2 (left), messages such as "because he/she looks guilty" or "because he/she acts suspicious" could, for example, be constructed.

5 User Study

To test the possible benefit of a pictorial interaction language over traditional interaction modalities, a user study was conducted with 9-11 year olds.

5.1 Interaction Modes

For the present study, we implemented two interactive versions of our system both using the touch-based interaction on the iPad: icon-based vs. menu-based. The icon-based version contains the pictorial language described above. The menu-based version was implemented to provide a set of choices in text form, representing choices that could also be constructed with the pictorial interaction language, which can be selected by the user by clicking on them (see Figure 3 for comparison of the two iPad interfaces).

The setup of the game is constant for both versions. In each case the friend character repeats the choice of the user, comments on it and returns to the group. However, the characters' comments are limited to the number of choices in the menu-based version to ensure that users are not influenced by the wider variety of the system in the icon-based version.

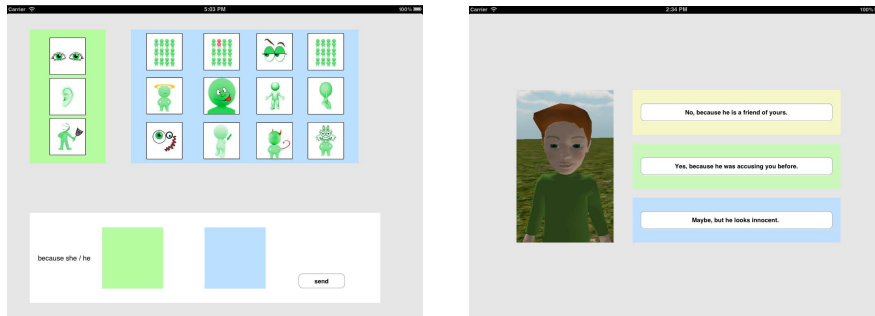


Fig. 3. Screenshots of the iPad showing icon-based interaction (left) and menu-based interaction (right)

5.2 Expectations

With the pictorial interaction language, we provide interpretational freedom to the users by offering many opportunities to construct sentences. In addition we want to inspire children's curiosity and exploration. For our study, we hypothesized that an icon-based interaction style would be perceived as more engaging and interesting compared to a traditional menu-based interaction.

However, a possible advantage of the menu-based version might be that it is more intuitive to use for inexperienced users, since fewer, clearer choices are provided and there is no need for children to construct their own sentences.

5.3 Procedure

To investigate which interaction modality the children preferred, an evaluation study was conducted with the target age group with each child having a PC, iPad and headphones. Children were introduced to the study activities, before playing both versions of the system. After using each of the versions, the children

completed a questionnaire, then used the other version and completed the same questionnaire again.

For evaluation, a 4-part questionnaire was developed:

Part 1 provided requested descriptive data, e.g. the children's age, gender and previous exposure to tablets.

Part 2 included questions focused at gaining the child's response to their first interactive experience. Each question was provided as bi-polar adjectives using a 5-point facial scale, see Figure 4. Facial scales have been shown to be well suited for evaluation with children in school environments, see [16]. The questions included:

- Ease of use: was the application easy to use or not, and could the child achieve what they intended with the interaction
- Engagement: was the experience fun, was it exciting, would they want to play again, would they have liked to play longer
- Visual appeal and interaction comprehension: did children like the appearance of the interface and could they understand the meanings provided in the interaction dialogue (e.g. the menu items or the icons)
- Open questions asking what children liked most and least about the game

Part 3 repeated the questions in Part 2 for the second interaction experience.

Part 4 asked the child to compare the two interaction modalities and decide which had been more fun, exciting and interesting.

Finally children were given a gold star sticker and were asked to put the sticker on a picture of the version they liked the best. The gold star sticker was chosen as children recognise stars and stickers as tokens that are awarded for something that is very good i.e. stars and stickers are often given in class for a good piece of work.











Do you think that the Werewolves game on the iPad was:						
Easy to use						Not easy to use
Fun						Boring

Fig. 4. Example questions from the questionnaire

71 children aged 9-11 years ($M = 9.65$, $SD: .56$) living in the UK participated in the study. 59.2% ($n = 42$) of the sample was boys, and 40.8% girls ($n = 29$). Most children had used an iPad before (84.5%, $n = 60$).

35 of the children used the icon-based application first, followed by the menu-based version, while the remaining 36 children played the versions ordered the other way round (i.e. the procedure was counterbalanced to avoid order or practice effects).

6 Results

Mean values of the children’s ratings are summarized in Figure 5. It shows that overall the icon-based version was rated more positively compared to the menu-based version. There was one exception to this for ratings of the pictures on the iPad being easy to understand / hard to understand. For this question, children rated the menu-based interaction slightly more favourably (i.e. the menus were easier to understand).

Figure 5 also clearly illustrates that children rated all the questions positively for both the menu and icon-based interaction, with no mean ratings above 3 (scale ranged from 1 to 5, with 1 being most favourable and 5 the least favourable). The highest (i.e. least favourable) mean response of 2.61 was for ratings of how exciting / dull the menu-based interaction was.

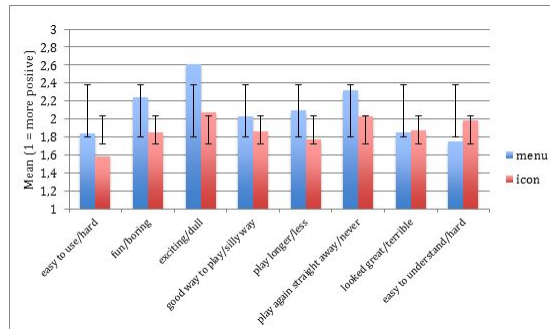


Fig. 5. Mean ratings (error bars 1 SD) of icon-based vs. menu-based interaction

The facial scale that children rated could not be assumed to follow an interval scale, but rather ordinal. Therefore, non-parametric Wilcoxon signed-rank tests for related samples were calculated to determine whether there were any significant differences in mean ranks between children’s ratings of the menu-based versus icon-based interactions. Figure 6 illustrates the Z statistic, associated

Question	Z value	p (1-tailed)	r
Easy to use/not easy to use	-1.92	.03	.23
Fun/Boring	-3.10	.001	.37
Exciting/Dull	-3.66	<.001	.43
Good way to play/silly way	-1.07	.14	.13
Play longer/less time	-2.31	.01	.27
Play again straight away/not at all	2.14	.02	.25
looked great/terrible	-.03	.49	.00
easy to understand /hard to understand	-1.52	.06	.18

Fig. 6. Interaction questions for the menu-based and icon-based interaction, associated Z values, significance (1-tailed), and effect size (r)

significance for one-tailed tests, and effect sizes (r) for each of the questions in our questionnaire.

Children rated the icon-based version (median = 1) as slightly easier to use than the menu-based interaction (median = 1). The icon-based interaction (median = 2) was also rated as more fun to use compared to the menu-based interaction (median = 2). Children rated that the icon-based interaction (median = 2) was more exciting compared to the menu-based interaction (median = 3) and wanted to play longer with the icon-based (median = 1) interaction compared to the menu-based (median = 2) interaction. The icon-based interaction (median = 2) was rated more favourably by children for wanting to play again straight away compared to the menu-based interaction (median = 2).

No significant differences were found between the icon (median = 2) and menu-based (median = 2) interactions for ratings of whether it was a good way to play the game. Children rated the design of the interface (looked great/looked terrible) on both the menu-based (median = 2) and icon-based interaction (median = 2) favourably, with no significant differences reported between the two interactions. Children also found the interface for both the menu-based (median = 1) and icon-based (median = 1) version easy to understand, with no significant differences. Figure 7 illustrates the positive (icon ζ menu) and negative (icon \jmath menu) ranks derived from the Wilcoxin tests for each of the eight questions that children rated using the facial scale. The figure clearly demonstrates that children favoured the icon-based interaction over the menu-based interaction for all questions, with the exception of whether the interaction interface on the iPad was easy/hard to understand.

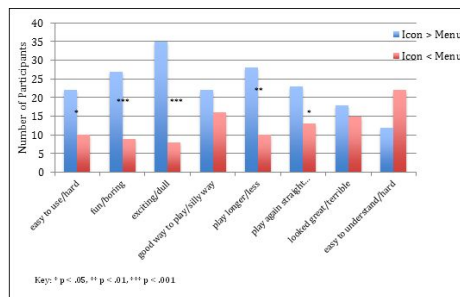


Fig. 7. Number of participants who rated the icon-based interaction more favourably than the menu-based interaction (icon ζ menu), and the menu-based interaction more favourably than the icon-based interaction (icon \jmath menu)

Binomial tests (0.50) were carried out to determine whether children found the icon-based or menu-based interaction more fun, exciting and interesting. Children reported finding the icon version more fun compared to the menu version [fun ($Z = 5.93$, $p \jmath .001$), menus $n = 10$ (.14), icons $n = 61$ (.86)], and the icon version more exciting [exciting ($Z = -2.85$, $p \jmath .01$), menus $n = 23$ (.32), icons $n = 48$ (.68)]. No preference for menus or icons was reported by children

for levels of interest [interesting ($Z = .000$, $p = 1.0$), menus $n = 36$ (.51), icons $n = 35$ (.49)].

Each child was given one sticker and asked to place this on his/her favourite version of the interaction interface. 92% ($n = 55$) of children placed their sticker on the icon-based version, leaving just 8% ($n = 5$) of children who placed it on the menu-based version. A one-sample binomial test revealed that significantly more children said that the icon-based version was their favourite compared to the menu-based version [favourite ($Z = 6.33$, $p < .001$), words $n = 5$ (.08), pictures $n = 55$ (.92)].

7 Conclusions and Future Work

This paper discusses the development and evaluation of a pictorial interaction language to test the suitability of such an interaction modality for 9-11 year old children for a cultural learning scenario. We compared the experience of the pictorial interaction language with a more traditional menu-driven interaction. Through using the same application with the same interaction device we have established that children preferred the pictorial interaction, considering it to be more fun and exciting than a menu-driven approach. In line with our expectations, the children rated the pictorial interaction as harder to understand compared to the menu-driven approach but at the same time more fun. We thus think that the pictorial language provides a more challenging interaction that positively influences the overall user experience. Our focus was to investigate the potential for a pictorial interaction approach to engage children in a games-based learning experience. Through directly comparing pictorial interaction and menu-driven interfaces, even though children were positive about both approaches, results indicate that:

- Children found the pictorial interaction to be more fun than the menu-driven version.
- Pictorial interaction was perceived as more exciting than menus
- Children would have liked to play longer with the pictorial interaction than with the menu-driven system
- Pictorial interaction is well suited for, and preferred by, the age group with just 8% of the children preferring the menu-driven version.

The pictorial interaction language is now integrated into the complete cultural conflict learning experience. Studies conducted in Germany and UK with the full system will further establish the benefits of an intuitive pictorial interaction language for supporting children in developing cultural understanding and awareness.

Currently the whole system is prepared for usage with Japanese children, to investigate whether the pictorial interaction language and the serious game are understood in an Asian culture as well. Therefore, the provided grammatical structure of the interface had to be slightly adapted.

Acknowledgments. This work was funded by the European Commission within the 7th Framework Program under grant agreement eCute (education in cultural understanding, technologically enhanced).

References

1. Bowman, D.A.: 3D User Interfaces. In: The Encyclopedia of Human-Computer Interaction. The Interaction Design Foundation, Aarhus (2013)
2. Forlines, C., Wigdor, D., Shen, C., Balakrishnan, R.: Direct-touch vs. mouse input for tabletop displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2007 (2007)
3. Nazir, A., Ritter, C., Aylett, R., Krumhuber, E., Swiderska, A., Degens, N., Endrass, B., Hume, C., Hodgson, J., Mascarenhas, S.: ECUTE: DIFFERENCE IS GOOD! In: IADIS International Conference on e-Learning, pp. 425–429 (2012)
4. Aylett, R., Lim, M.Y., Hall, L., Endrass, B., Tazzyman, S., Ritter, C., Nazir, A., Paiva, A., Hofstede, G.J., Andre, E., Kappas, A.: Werewolves, cheats, and cultural sensitivity. In: Proc. of 13th Int. Conf. on Autonomous Agents and Multiagent Systems, AAMAS 2014 (2014)
5. Plotkin, A.: Werewolf: A Mind Game (2010), <http://www.eblong.com/zarf/werewolf.html>
6. Sali, S., Wardrip-Fruin, N., Dow, S., Mateas, M., Kurniawan, S., Reed, A.A., Liu, R.: Playing with words: From intuition to evaluation of game dialogue interfaces. In: Proceedings of the Fifth International Conference on the Foundations of Digital Games, FDG 2010, pp. 179–186. ACM, New York (2010)
7. Aylett, R.S., Vala, M., Sequeira, P., Paiva, A.C.R.: Fearnot! - An emergent narrative approach to virtual dramas for anti-bullying education. In: Cavazza, M., Donikian, S. (eds.) ICVS-VirtStory 2007. LNCS, vol. 4871, pp. 202–205. Springer, Heidelberg (2007)
8. Bondy, A.S., Frost, L.A.: The picture exchange communication system. Focus on Autism and Other Developmental Disabilities 9(4) (1994)
9. Mirenda, P.: Toward functional augmentative and alternative communication for students with autism. Language, Speech, and Hearing Services in Schools 34, 203–216 (2003)
10. Symbols Worldwide Ltd T/A Widgit Software: Widgit, <http://www.widgit.com/index.php> (last viewed: April 2, 2014)
11. SymbolWorld, <http://www.symbolworld.org/> (last viewed: April 2, 2014)
12. Antle, A.: The design of cbc4kids storybuilder. In: Interaction Design and Children (IDC 2003), pp. 59–68 (2003)
13. Kelleher, C., Pausch, R., Kiesler, S.: Storytelling Alice Motivates Middle School Girls to Learn Computer Programming. In: CHI 2007, pp. 1455–464. ACM (2007)
14. Takasaki, T., Mori, Y.: Design and Development of a Pictogram Communication System for Children Around the World. In: Ishida, T., R. Fussell, S., T. J. M. Vossen, P. (eds.) IWIC 2007. LNCS, vol. 4568, pp. 193–206. Springer, Heidelberg (2007)
15. Damian, I., Endrass, B., Huber, P., Bee, N., André, E.: Individualized Agent Interactions. In: Allbeck, J.M., Faloutsos, P. (eds.) MIG 2011. LNCS, vol. 7060, pp. 15–26. Springer, Heidelberg (2011)
16. Read, J.: Validating the Fun Toolkit: An instrument for measuring childrens opinions of technology. In: Cognition Technology and Work (2007)

Why Numbers, Invites and Visits are not Enough: Evaluating the User Experience in Social Eco-Systems

Lynne Hall, Colette Hume
Department of Computing, Engineering and Technology
University of Sunderland
Sunderland, SR6 0DD, United Kingdom
{lynne.hall, colette.hume} @sunderland.ac.uk

Abstract — Social eco-systems are often evaluated through quantitative data that is automatically logged and analysed. However, where the user’s experience of social eco-systems is evaluated, more explicit intervention approaches are typical, with questionnaires, focus groups and user testing widely used, directly asking the user about their experience. User experience evaluation thus ruptures the social eco-system, occurring as a separate, discrete activity outside of that system. In this paper, we propose that evaluation should be part of the social eco-system adding value to the user experience. We outline an evaluation approach that has been applied within games-based learning environments where the evaluation is seamlessly embedded. We briefly outline our approach to generating and analyzing data highlighting its potential for social eco-system evaluation.

Keywords—*evaluation; user experience; analysis of user-generated content.*

I. INTRODUCTION

It is widely recognized that the impact of social eco-systems requires further consideration, with relatively few studies or empirical investigations. Social eco-systems typically involve enjoyable and often affective interactions within a user-chosen context. The user’s interaction focus is primarily recreation, enjoyment or problem solving in relation to a social need. Yet, how can we evaluate or understand the impact of that interaction on an individual or societal level? And more, if we do try to evaluate it, can we do this without having an experimenter effect, even if that “experimenter” is an anonymous on-line survey.

Users are only vaguely aware and in general don’t seem to care about the collection of usage statistics. Thus, statistics can be endlessly calculated relating to the number and frequency of visits, invites, postings and so on, without any impact on the user. However, evaluating the user experience is more challenging, requiring conscious user input, rather than logging of actions.

Unlike the integrated usage data collection, the user experience evaluation of social eco-systems is typically a separate, discrete activity to the main use of the system, with questionnaires, focus groups and user testing widely used. User experience evaluation thus changes the dynamic of the

social eco-system, placing the user in the role of evaluator rather than social network member.

In this paper, we propose an alternative to this discrete, separate approach to user experience evaluation. Instead of separating out evaluation and changing the role of the user, we have developed an approach that enables us to evaluate the user experience without users being aware that they are taking part in an evaluation. This approach has considerable relevance to the evaluation of social eco-systems, meeting two key success factors for social networks:

- Evaluation should be invisible and should have no (as achieved with usage statistics) or a positive impact on user activities
- Add-ons (e.g. evaluation instruments) to the social eco-system must be integrated and add value to the user experience

In this paper, we briefly outline our approach to the generation of evaluation content and discuss our proposed approach to the analysis of this content. Our key focus is how to mask the evaluation experience so that the user is unaware of their evaluation input whilst generating data useful to an interdisciplinary research and design team. This approach has been successfully applied and we believe that it offers potential for other developers and researchers to evaluate social eco-systems. Section 2 briefly discusses social eco-system evaluation, highlighting the focus on commercial factors and the relevance of these to user experience evaluation. Section 3 discusses our approach to user experience evaluation, outlining our approach and its application to two systems. Section 4 discusses our approach and considers its potential for evaluating social eco-systems. Section 5 concludes that this approach has considerable relevance to supporting and improving the user experience of evaluation.

II. EVALUATING SOCIAL ECO-SYSTEMS

There has been a massive growth in commercially supported social eco-systems. The marketers, quite rightly, recognize that supporting an on-line community will increase brand loyalty and sales. Through allocating significant resources to on-line activity, some companies have established high quality, effective social eco-systems, with significant user presence. The purpose of these social eco-systems is to enable companies to achieve their business

goals. Thus, in the evaluation of such commercially derived social eco-systems the evaluation issue is not really user experience and social impact, rather it is the company's Return On Investment (ROI). This ROI includes the social eco-systems impact on: developing brand loyalty, thought leadership, reducing operating costs, optimizing marketing budgets, and increasing profits [1].

With the aim of demonstrating ROI, much of the evaluation in social eco-systems is achieved using logged user interactions. For example, the number of invites made by a user; frequency of postings; and number and type of interactions within the social eco-system. There are many tools available to log and analyse user interactions, with such functionalities increasingly provided as standard in site development products. However, whilst tools can be used as a basis to calculate a range of quantitative measures such as visits, social graph, social surface area, etc. their insight into the direct user experience is limited. Whilst such numerical data can enable us to determine the strength, sustainability and growth potential of the social eco-system, it does not allow us to explore the user experience itself.

There are considerable challenges for user experience evaluation of social eco-systems, with users often geographically dispersed and having limited real world interactions. In response to this, techniques have been developed for both virtual and real world evaluations. However, the majority of these require additional user input, often with the user role changing from member, player, commentator, etc. to a critic, tester or evaluator.

Whilst engagement in user experience evaluation can offer positive benefits to participants, for example, early access to new features, input to development, status within the network, etc., many users choose not to participate in evaluations. Thus, unless participation in evaluation activities is mandated (e.g. in a fiat system [2]), the participants self-select thus providing only a partial view of the user experience of the social eco-system. Further, where participation in evaluation activities is mandated, users can view evaluation as a burden [3].

In considering the evaluation of the user experience in a social eco-system, it is not the issue of usability that is key. There are a whole variety of half-hearted attempts by companies and organisations to create social eco-systems. From these, we know that if the usability is poor that unless the environment is incredibly compelling, then users will go elsewhere. Instead, it is the user's personal, social and emotional experience that requires evaluation to enable us to explore the impact of social eco-systems.

III. EMBEDDING EVALUATION IN THE USER EXPERIENCE OF SOCIAL ECO-SYSTEMS

Our approach to evaluation has been developed within the EU FP6 eCIRCUS [4] and FP7 eCUTE [5]. Both projects have focused on technology enhanced learning for significant social issues, including bullying and intercultural conflict. In this paper, we discuss our evaluations with the ORIENT [5] and MIXER [6] showcases, outlining our

approach and highlighting the potential for its use with other social eco-systems.

Our research has focused on evaluating a specific type of social eco-system: technology enhanced learning through interaction in intelligent computer assisted role-play environments. In our experiences of designing, developing and evaluating our showcase applications, we have dramatically changed our approach to evaluation. Rather than evaluation being conducted as a discrete, separate activity to the interaction, we now add value through seamlessly embedding evaluation into the user experience. The impact of this is that users are unaware they are taking part in an evaluation. In addition, the results from this evaluation have been of considerable use to the interdisciplinary development team.

To enable us to evaluate our showcases, users are actively engaged in the individual and communal generation of real world artefacts and digital assets. Critical to the success of our approach is for users to be aware of, and participate in the social eco-system provided through our environment. We artificially create a temporary social eco-system for a specific showcase and its participating users. Whilst we have to stimulate users into creating assets, in many social eco-systems a plethora of such user-generated content exists or could easily and enjoyably be developed meeting the requirements of the evaluation and improving the user experience.

However, having extensive data or content is insufficient without a viable analysis approach. Analysing the content is complicated by a multiplicity of formats and the challenges offered by non-textual assets. Our evaluation approach uses a range of techniques and tools for content analysis, with approaches derived from information retrieval research transforming the content into usable data.

The following examples briefly outline our approach to generating and analyzing user experience data.

A. *ORIENT: Seamlessly embedding evaluation into the user experience*

ORIENT provides users with an intelligent computer assisted, semi-immersive, graphical role play environment depicting an imaginary culture, the 'Sprytes.' It is aimed at teenagers and young adults who interact in groups of 3, taking roles in Space Command (a benevolent United Nations type of organization with a galactic focus) with the goal of helping the Sprytes to save their planet from imminent destruction. ORIENT's learning focus is cultural understanding and sensitivity.

The characters, the Sprytes, inhabiting this world are autonomous agents, based on an extension of the FATiMA agent architecture [7]. Emotional appraisal is based on the OCC cognitive theory of emotions [8] extended by incorporating aspects of a needs driven architecture, PSI [9]. To enable cultural adaptation of the agents, Hofstede's cultural dimension values were added to the agent minds for the culture of the character; cultural specific symbols; culturally specific goals and needs, and the rituals of the culture [10].

Users interact with the Sprytes using a Wiimote to provide gestures and speech recognition of character names. They interact with the ORIENT world using a scanner phone with the ORA-CLE (Onboard Resource Agent - Cultural and Liaison Engagement), a mobile phone based embodied conversational agent whose role is to support the users in their interaction. Figure 1 provides an overview of ORIENT's main components. At the core of the system is the virtual world model that is presented to the user as 3D graphics on a large screen, in front of which the users interact with ORIENT as a group.

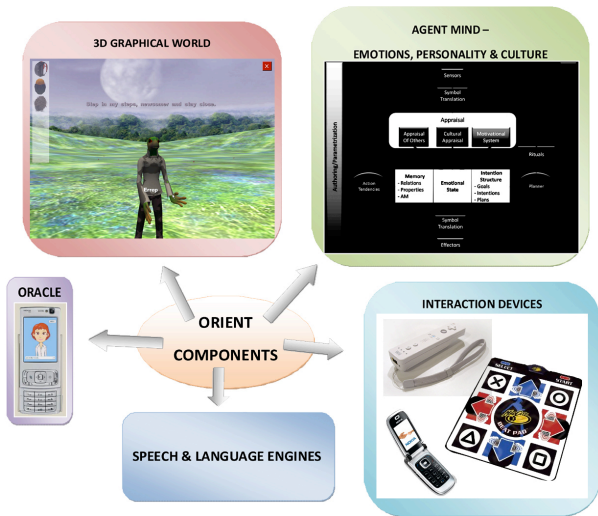


Figure 1. ORIENT Overview

Developed as part of an interdisciplinary project, the evaluation aimed to investigate the effectiveness of ORIENT in fostering cross-cultural acceptance through the promotion of collaborative practices and the appreciation of similarities and differences between cultures. From the technical perspective, evaluation focused on the coherence and comprehensibility of the narrative; the believability and credibility of the agents that underpin the characters; and participant engagement with the cultures of ORIENT and the Sprytes themselves. With the interaction approach, we focused on evaluating the participant's views of the impact of unusual interaction devices and mechanisms, focusing on device usability and user satisfaction with unusual interaction mechanisms. This resulted in a wide range of purposes and instruments required for the evaluation.

Even though we needed users to participate in an extensive evaluation, our goal was for players to have only one consistent experience that of being a player in a role play game. To achieve this we transformed traditional and/or well established data gathering instruments into 'in role' counterparts. These were then embedded into the role play and reinforced with supporting artifacts. Each instrument was given archetypal branding (adding value to the role play context) and an age appropriate format and aesthetic (meeting user expectations), see figure 2. The resulting battery of piloted instruments aimed to add maximum value

to the over-arching role playing game while collecting key evaluation data to help developers assess the user experience from a number of theoretical perspectives.

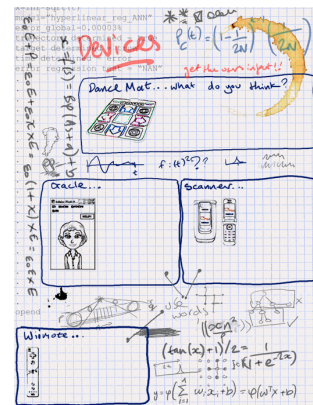


Figure 2. Evaluation Instruments

This approach was very successful in generating data from users about their experience, with the interesting side-effect that users were completely unaware where the game stopped and the evaluation started. The evaluation instruments and activities are effectively seamless and thus data captured in a way that is invisible for the user. Rather than the evaluation instruments and supporting artifacts adding a burden to the user, they seemed instead to enhance the game, actually increasing the immersion and enjoyment of the users. The data and content produced through the user interactions was analysed using qualitative and quantitative analysis techniques and are further discussed in [11].

B. MIXER: Opinion and sentiment: approaches to analyzing user generated content

With ORIENT, the majority of the user-generated content was achieved through specially prepared instruments many of which were hard copy. With the ongoing development of our evaluation approach, we are focusing on the generation of digital assets. Our exploration of the generation and analysis of digital assets is currently focused on MIXER [6]. This application aims to provide 9-11 year olds with classroom-based, technology enhanced learning experiences related to cultural conflict. This context for MIXER is provided by Hide & Seek where participants may be characters or other users and where conflict is typically a result of rule misunderstandings, based on Hofstede's cultural dimensions [10]. Figure 3 provides some frames outlining the MIXER narrative.

Our evaluation is focused both at children and teachers as achieved through their interactions with MIXER and their discussing of these experiences. The evaluation is seamlessly embedded into the experience of the application, right from the initial design stage. For example, the frames in figure 3 have been generated as a comic book. Into this comic book (which represents the application for the users) we have embedded traditional questionnaires that have been morphed into quizzes and mini-games.

In addition, the comic book is supplemented by an on-line experience, where the users will engage in the generation of blogs, digital AV & photo albums and participation in a tailored social network. Two complementary social networks are used, one for the teachers and the other for the child users of MIXER.

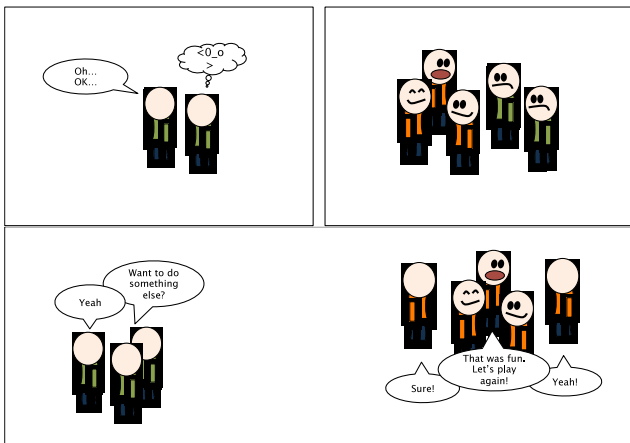


Figure 3: Frames from the MIXER storyboard

During the user interaction, a main focus is gaining an appreciation of the user's Theory of Mind with users prompted with a range of questions which result in the user producing freeform text in the interaction. In addition to this direct input during the interaction users are also involved in generating content in relation to their experience.

An initial study, involving three groups of 9-11 year old children (10 in each group) has recently been conducted to establish user responses to embedding evaluation in MIXER.

The three conditions were provided in separate locations. This separation was to ensure that children of one group were not aware of the other conditions' activities. The three evaluation conditions were:

- **Direct evaluation:** children were provided with a non-interactive comic book (just for reading); a work book composed of a series of activities related to the comic book and specifically activities related to Theory of Mind; and a set of questionnaires related to the comic book, attitudes to culture, in-group / out-group, understanding of cultural dimensions
- **Hybrid evaluation:** children were provided with an interactive comic book and asked to write / draw responses (thus incorporating Theory of Mind activities). The workbook (without the Theory of Mind activities) and questionnaires were given as a single item, with the questionnaires embedded in the workbook.
- **Seamless evaluation:** children were provided with a single artefact incorporating the interactive comic book, the workbook activities and the questionnaires. The questionnaires were modified and presented as quizzes and activities using age appropriate aesthetics. The activities and questionnaires were placed throughout the comic book, replicating the approach of magazines for 9-11 year olds.

Observations during the use of MIXER highlighted that children in the Seamless Evaluation condition engaged for longer and were highly engrossed in the workbook and evaluation activities. Children in this condition required very little input or encouragement from the adults present and worked steadily through the entire artefact. More questions and issues were raised in the other conditions, particularly in relation to completion of the questionnaires. In the direct evaluation condition, children were not particularly interested in completing the questionnaires and spent significantly less time with the comic book and Theory of Mind activities, than those in the Hybrid or Seamless Evaluation conditions.

We are currently engaged in analyzing the data generated during this initial MIXER study. Early results indicate that the results from using different versions of the questionnaires are relatively similar across the conditions. This is an expected outcome illustrating that although instruments are modified they are essentially collecting the same data. However, in line with their greater engagement, children in the Seamless Evaluation condition wrote and drew more within the comic book (and the embedded Theory of Mind activities) than the other conditions. Our initial results appear to indicate that improving the user experience of evaluation results in greater user engagement. Current work focuses on further analyzing our data, particularly in relation to the impact of embedding the Theory of Mind within the Comic book.

With MIXER, we are now focusing on the analysis of freeform text and digital (e.g., audio, video, photos)

contributions. Our key aim in evaluating these user-generated content is to determine personal, social and emotional user experience. As such, we are particularly looking for opinions and affective views within user-generated content.

There is current considerable interest in the evaluation of “opinionated content” such as discussion groups, blogs, tweets, video postings and other methods where people express their views online. Through evaluating relevant user-generated content, it is obvious that companies can gain consumer feedback about their own and competitor’s products, thus avoiding the need to conduct surveys, organise focus groups or employ external consultants [12].

A considerable number of statistical measures can help analyse text and automation tools. Through the use of semi-automated methods, we will be analyzing user generated content provided in the MIXER social eco-system. We are currently investigating a range of methods, aiming to find the most appropriate analysis approach for our evaluation purposes. These interdisciplinary evaluation purposes are quite broad, relating to educational, psychological, socio-cultural, interaction and technical goals. Methods we are investigating include:

- Use of base and comparative polar words (e.g., base: “bad”, comparative: “worse”) enabling the use of statistical measures (e.g., [13]).
- Seed words and connectives such as AND, OR, BUT, or HOWEVER are being used to find related or contrasting words, as in [14].
- Clustering techniques, such as Factor Analysis are being used to identify word and opinion clusters.
- Named entity recognition (as applied in ontology generation) will be applied, aiming to support co-reference resolution, for example – a pronoun such as “it” might refer to “the game”, “MIXER”, “the computer,” etc.
- Synonym grouping will be facilitated using SentiWordNet (as used by [15])

A key benefit of using sentiment analysis is that it can be used to convert natural language texts into structured data, that can then be stored and manipulated in a database. We will use Liu’s approach [12] and store user generated content as a quintuple:

- Object (product, person, event, organisation, topic),
- Feature of that object
- Polarity of the opinion of the holder on that feature of that object
- Opinion holder
- Time when the opinion made by opinion holder

This data can be both analysed statistically and represented visually, supporting a greater understanding of the data. Although we have just begun applying this approach to our analysis of MIXER, early investigations suggest that this will provide a powerful addition to our evaluation approach.

IV. DISCUSSION

The Six Benchmarks for Digital Marketing Strategy [16] have been developed to evaluate the potential effectiveness of social media on ROI:

1. Goal - What is the targeted goal of your advertisement, social media program or campaign?
2. Engage – How effective is the message in attracting or involving your target market?
3. Relationship – Did the message stimulate the target to feel trust or common interests?
4. Value –Does the product or service and related message communicate added benefit for the individual, organization or company?
5. Action- Does the message move you to act?
6. Synergize- Is the tool an add-on to current marketing efforts or is it integrated into the campaign?

Although such benchmarks identify plentiful questions and issues, there is little information about how systems can be evaluated against them. Whilst usage stats will answer some issues, clearly, user experience data has to be both generated and analysed to permit evaluation against these benchmarks.

In this paper, we have proposed an approach to the generation and analysis of user generated content. Our approach differs from many current user experience evaluation approaches. Through focusing both on reducing the visibility of evaluation participation and on adding value through evaluation our approach gains useful data whilst either having no or a positive affect on the user.

Our approach to gathering user experience data involves the use of existing user input formats (e.g. blogs, postings, tweets) and the creation of add-ons (e.g., questionnaires represented as quizzes, mini-games, etc.). Our users are consistently unaware that they are taking part in an evaluation. Results have highlighted that users view the evaluation experience positively, seeing it as a value add rather than a negative. In addition, the interdisciplinary project team have gained results and evaluation data that have been relevant and useful.

Within our approach, we are gathering data in two ways. Firstly, through crafting customized quizzes and embedding questions (from existing traditional questionnaires) in interactions and entertaining activities. And secondly, through viewing user generated content as a primary source of evaluation material. Where possible we avoid technology learning and thus use popular formats, Facebook has already trained most of our users.

Sentiment analysis and opinion mining offer considerable potential for the analysis of user generated content in the evaluation of any social eco-system. Semi-automated approaches can greatly increase the speed of data refinement and analysis. The use of such approaches also provides the data in a format that is relatively easy to visualize, thus allowing greater understanding by development teams and stakeholders.

Related work focuses on the evaluation of AV and photographic content. With photography we are exploring indexicality to support evaluation [17]. With both photographs and AV content, the critical issue is how to transform the content into analyzable outputs. Initial results suggest that the labels and descriptions frequently generated by users along with non-textual postings may contain sufficient content to analyse the AV without requiring additional data refinement. To further investigate we are exploring the use of meta-tagging, to enable us to compare results from further content refinement with the use of user generated labels and descriptions.

V. CONCLUSIONS

It is possible to create a user experience evaluation that can be completely embedded within a social eco-system. Evaluation instruments and approaches can be crafted to enhance rather than detract from the social eco-system experience. Sentiment analysis and opinion mining transform user generated content into a highly valuable and analyzable data source. The use of this approach allows user experience evaluation data to be gained and analysed as invisibly as usage statistics.

VI. ACKNOWLEDGMENTS

This work was partially supported by European Community (EC) and is currently funded by the ECUTE project (ICT-5-4.2-257666). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

VII. REFERENCES

- [1] Kaufman, I., (2010) 5 Steps to Evaluation Social Media ROI, Social Media Today, Available at: <http://socialmediatoday.com/irakaufman1/111688/5-steps-evaluating-social-media-roi> Accessed 25-07-2011
- [2] Hinchcliffe, D., (2010) The K-factor Lesson: How Social Ecosystems Grow (Or Not), Enterprise Irregulars, Available at: <http://www.enterpriseirregulars.com/9499/the-k-factor-lesson-how-social-ecosystems-grow-or-not/> Accessed 25-07-2011
- [3] Sapouna, M., Wolke, D., Vannini, N., Watson, S., Woods, S., Schneider, W., Enz, S., Hall, L., Paiva, A. and Aylett, R. (2010) Virtual Learning Intervention to Reduce Bullying Victimization in Primary School: A Controlled Trial, *Journal of Child Psychology and Psychiatry*, 51(1), pp. 104-112.
- [4] <http://www.e-circus.org> Accessed 25-07-2011
- [5] Hall, L., Jones, S., Aylett, R., André, E., Paiva, A., Hofstede, G.-J., Kappas, A., Nakano, Y., and Nishida, T. (2011) Fostering Empathic Behaviour In Children And Young People: Interaction With Intelligent Characters Embodying Culturally Specific Behaviour In Virtual World Simulations, in *INTED 2011 (International Technology, Education and Development Conference)*, Valencia, Spain, 7-9 March, 2011.
- [6] Hall, L., Hall, M., Hodgson, J., Nazir, A. and Lutfi, S. Games based learning for Exploring Cultural Conflict, in *AISB 2011 Symposium: AI & Games*, York, 6-7 April, 2011.
- [7] Dias, J., and Paiva A. (2005) *Feeling and Reasoning: a Computational Model*. 12th Portuguese Conference on Artificial Intelligence, EPIA (2005), pp. 127-140.
- [8] Ortony, A., Clore, G. L., and Collins, A. (1988) *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, UK.
- [9] Dörner, D. (2003). The mathematics of emotion. *The Logic of Cognitive Systems: Proceedings of the Fifth International Conference on Cognitive Modeling*, Springer, pp. 127-140.
- [10] Hofstede, G., Hofstede, G.J. and Minkov, M. (2010) *Cultures and Organisations, Software of the Mind, Intercultural Cooperation and It's Importance for Survival (3rd Edition)* McGraw-Hill Books, New York
- [11] Aylett, R., Paiva, A., Vannini, N., Enz, S., Andre, E. and Hall, L. (2009) But that was in another country: agents and intercultural empathy, 8th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, pp. 329-336
- [12] Liu, B. (In press). Sentiment Analysis and Subjectivity. To appear in *Handbook of Natural Language Processing*, Second Edition, (editors N Indurkha and F Damerou).
- [13] Turney, P. (2002). Thumbs up or thumbs down. *Semantic orientation applied to unsupervised classification of reviews*. *Proc. ACL 2002*, pp. 417-424.
- [14] Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. *Proc. Joint ACL/EACL conference*, pp. 174-181.
- [15] Devitt, A. and Ahmad, K. (2007). Sentiment Analysis in financial news: a cohesion-based approach. *Proceedings of ACL*, pp. 984-991.
- [16] Kaufman, I. (2011) Six Benchmarks for Digital Marketing Strategy, Social Media Today, Available at: <http://socialmediatoday.com/patsystemart/262270/gervas-6-benchmarks-digital-marketing-strategy> Accessed 25-07-2011
- [17] Jones, S. and Hall, L. (2011) The photograph as a Cultural Arbitrator in the Design of Virtual Learning Environments for Personal and Social Education for Children and Young People. *Electronic Visualisation and the Arts 2011*, London, 6-8 July, 2011.

SCAFFOLDING THE STORY CREATION PROCESS

Hall, M., Hall, L., Hodgson, J., Hume, C. and Humphries, L.

*Department of Computing, Engineering and Technology, University of Sunderland, St. Peters Way, Sunderland, UK
{marc.hall, lynne.hall, john.hodgson, colette.hume, lynne.humphries}@sunderland.ac.uk*

Keywords: Comic books: Technology: Education: Prototyping.

Abstract: Comic books provide an appropriate and structured context for education and personal or peer reflection. In this paper we discuss the benefits of comic books and technology in a pedagogical context, including the mechanism of scaffolding and how this interaction impacts upon the child's environment. Our studies into the educational benefits of comic books have led to the development of an interactive comic book application. The application is being developed for the purpose of narrative inquiry through the creation and completion of a story scaffold. The analysis of the data will help evaluate the child's social and cultural interaction with the story

1 INTRODUCTION

In comparison to traditional textual narration, it has been seen that the process of completing stories partially defined as comic books or graphical novels provides an appropriate and structured context for eliciting affective and reflective thought (West et al., 2004; Pennington et al., 2011). Comic books have been used as an engaging and motivational learning activity for both adults and children (Norton, 2003). They are appropriate for the classroom (McVicker, 2007), encouraging the development of critical thinking skills (Birisci & Metin, 2010.)

Whilst some research implies that the use of comics in the classroom is most applicable specifically to male students with low attainment levels, there is also some evidence that they can also be used successfully with high achievers (Sabeti, 2011; Lenters, 2006; McTaggart, 2005; Norton, 2003; Cleaver, 2008). Comics have been used in a range of educational contexts, from Primary School through to University level. They have been used in developing understanding of concepts such as Business Ethics (Gerde & Foster, 2007), logic in Computer Science (Cervesato, 2011), science lab safety (Di Raddo, 2006), collaboratively teaching English as a Second Language (Sachs et al., 2003) and teacher education (Herbst et al., 2011).

There is wide use of technology enhanced learning, with applications ranging from the use of multimedia through to mobile devices (Stelzer et al., 2008; Ruchter et al., 2010). The use of technology has been found to be of particular benefit in supporting and developing literacy skills through the

use of new practices (Burnett et al., 2006). Such practices include peer-based learning activities such as groups of children sending emails in which each participant adds another line of the story to build up a collaboratively written narrative (Figa, 2007); and the use of PowerPoint in allowing the choice of appropriate images and text to help the learner to consider their audience (Abas & Zaman, 2010; Robin, 2008).

The use of comics as a pedagogical instrument removes some of the typical workload involved in story creation. It allows users to draw upon familiar presentation and scene structuring paradigms learned from prior experiences with comics, whilst having a positive impact on motivation (Bitz, 2008; Pelton et al., 2007). Furthermore, the intuitive nature of the comic book style makes it easy to learn for those that lack experience with the medium. Illustrated texts offer a unique perception of the narrative provided to the reader and have been shown to create a more empathic sense, allowing more evaluative and critical responses (Moschovaki & Meadows, 2005; Williams, 2008).

The popularity of community driven comic creation by amateurs on the web has increased in recent years (Lopes et al., 2009). Sites such as Toondoo invite educators to create class accounts which allows for the sharing and peer review of completed stories. Such websites allow users to create simple narratives with predefined content, and also allow users to suggest the next stage in the story (Williams & Barry, 2005) in a linear manner.

Research by Jong (2009) found that in tasks in which information is presented as non-linear text (specifically hyper-link connected text) the increased

cognitive load of navigating the non-linear structure reduced user retention. This raises concerns about the possible negative impact of making a scenario non-linear. However, as Jong's work focused on information with no narrative or temporal progression it isn't clear if it is entirely applicable. It does seem to suggest that non-linearity should be applied with caution.

The aim of the interactive comic book application described in this paper is to draw upon the advantages of comic books in a pedagogical context. The interactive comic book application provides a scaffold around which users can create a completed story based upon their own experiences and understanding of the subject matter. This is portrayed within the comic in a constructive manner, and later allows self- and peer-reflection upon that content. Children's social interaction with artifacts are culturally mediated (Vygotsky, 1978) and although Vygotsky never used the term scaffolding, the use of comic narratives as a scaffold overcomes criticisms of the Vygotskian approach that it does not take into account a child's cognitive development.

The comic book provides a medium that is widely used and accessible to all children. Comics can match tasks to the child's zone of proximal development, bridging the gap between what the child can do without help and what they can do with help. With this application, the scaffold is provided as a scene graph: a conceptual node graph that defines all possible scenes and the choices that must be made to move between these scenes. Users simultaneously interact with this scene graph node-by-node, defining a specific scenario through that graph by the choices they make during the interaction; and fill in content to complete the story (for example by writing dialog or narration entries.)

2 PEDAGOGICAL CONTEXT

In order to give the scenario content a theoretically valid basis the groups of characters in the scenario and its plot were designed to reflect aspects of the cultural model proposed by Hofstede (Hofstede, 2010). The Hofstede model defines all cultures as a combination of five Cultural Dimensions: Power Distance, Identity, Gender, Uncertainty, and Virtue.

The use of these cultural dimensions is the basis for assessing the effectiveness of the comic as a pedagogical tool. To achieve this users were given a separate questionnaire instrument, the Inter-group Anxiety Scale (Stephan & Finlay, 1999) to measure

their level of inter-group cultural sensitivity and empathy directly. The IAS is a validated psychometric test that examines children's disposition toward out-groups, formalized as a level of anxiety.

The users were also asked to complete an 'interactive' comic. The content the users choose to add to the scenario, in response the inter-group situations presented in the story, would in effect be an indirect measure of the user's inter-group sensitivity and empathy. For example, the way the user chooses to have a character respond verbally to a situation or how they portray the character's state of mind reflect the user's inter-group sensitivity. By comparing the direct and indirect measures the efficacy of the comic book could be established. Further the general level of engagement with the process was assessed qualitatively.

A single cultural dimension was selected to simplify the task of implementing a scenario based on Hofstede's cultural dimensions. The selected dimension was Uncertainty Avoidance. This was incorporated into a story, entitled CampFire, which involves two groups who each manifest an extreme of the uncertainty avoidance dimension. One group focuses very much on the rules of play and micromanaging each other. The other focuses on a more carefree attitude where the rules mattered and group dynamics were important, but with more flexibility in how the game was played.

3 DEVELOPMENTAL STUDIES

Building on evidence in literature, that the use of comic books in a pedagogical context is itself effective, a preliminary study was conducted in order to validate the approach as a means to facilitate reflection on inter-group relations. For this study an 'interactive' comic book was used, which can be seen as a low-fidelity prototype of the final application, to establish that a comic book could elicit valid pedagogical impact.

3.1 Initial Study

The pilot study involved 2 groups of children aged 9-10. The children were given the comic book along side various activity sheets (containing the questionnaire instruments). The activity sheets were themed in the style of a childrens' activity magazine (rather than as sterile research instruments) and included a variety of activities based upon the subject matter of the narrative portrayed. The activity that was of particular importance to the

piloting of the comic book was the inclusion of the Inter-group Anxiety Scale (IAS).

3.1.1 Pilot One

The test group (20) received an interactive version of the CampFire comic in which speech bubbles and thought bubbles in the last pages of the comic were left blank. The comic was bound together with the IAS questionnaire and all the other activity sheets as a single workbook. A front cover and contents page were added and the documents were styled in a way that resembled a comic book annual.

2.1.2 Results

Results from the pilot study confirmed the comic book approach to be an effective means to engage children in inter-group reflection, and also an enjoyable and engaging experience for the children. When compared, the results of the two groups showed that the test group was able to comprehend the storyline of CampFire and add relevant and meaningful content to the storyline.

Further, the completion rate for the IAS in the test group was 85% compared to 10% of the IAS from the control group (all aspects of the workbook were intentionally optional so that level of engagement could be estimated). Qualitatively, the test group worked through and completed all activities contained in the annual requesting less help, where as the children in the control group struggled more with the activities.

In addition the pilot also provided results that had not been anticipated. The children provided more content than was expected or requested of them. For example, in both versions of the CampFire comic some of the faces of the characters were left blank, this was a design decision intended to enhance the minimalist look of the CampFire comic. In both groups the children drew in the missing faces to show emotions that were appropriate to the current scene.

3.2 Narrative Mode Study

In this experiment we aimed to identify whether children could complete a story based upon a story abstract concept and what the baseline for such an abstract story is. The problem is how to define a 'story' to an extent that participants have enough information to build a story but leave enough out of the definition such that the participants are being creative and not just adding to a predefined story.

3.2.1 Pilot Two

Users were put into groups of five and given a large sheet of paper on which a nine-tile empty comic strip was printed. The groups' task was to fill in these squares with illustrations, speech bubbles and thought bubbles (see Figure 1). The 'abstract story' was defined by considering a generic story arch about two friends who fall out, experience some important incident and then become friends again. This was defined and presented to the participants by, taking each box as a scene, specifying what the purpose of that box (or scene) is with respect to the story. For example, the purpose of the first box is 'to introduce the lead character.' Each of the nine boxes was given such a purpose, leaving the task of turning these abstract scenes into a specific story to the participants.



Figure 1: Shows a section from one of the large comics

The groups were encouraged to discuss and create a plan of what they would create. To do this they were given a small version of the empty boxes sheet onto which they could write short notes about what they would put in their final story. Once the plan was complete the groups were left to self-organise and complete the larger version comic in their groups.

3.2.2 Results

The results of the second pilot showed that the children found no difficulty in the task of completing an abstract storyline and grasped the concept of developing a story from the scaffold provided with ease. Each of the groups developed entirely unique storylines with coherent narratives that depicted a variety of experiences on subjects ranging from sport to damaging the environment to

Lady Gaga. The children were so highly engaged and enjoying the session that they complained when it was over. In the design of their comics the children also included conventional comic book visuals such as large red letters to visually express angry voices, without being prompted or advised to do so, supporting the principle that comic book are a natural and familiar environment for children

4 THE INTERACTIVE COMIC BOOK APPLICATION

For the purposes of clarity while describing the application it is necessary to define some key concepts. Familiar terminology will be used, but used in a way that is specific to this paper. The first concept is that of a *scene graph*: a network graph in which each node represents a scene and edges represent choices. This graph defines all possible scenes and choices available to the user and as such all possible scenarios, in a sense this could be said to encapsulate a meta-scenario. We will use the word scenario to refer to precisely one valid and complete path through the scene graph. The word narrative will be used to refer specifically to the content generated by the user (although the content the user provides isn't necessarily strictly narrative this word does seem to summarise roughly what the user creates.) The scenario graph provides the scaffold; the user assembles a scenario and 'fills in the gaps' with narrative. The interactive comic book application must provide two distinct but related functionalities to the two user groups, for easy distinction the user groups will be referred to as the *developers* and the *users*. The developer will use the application to construct an underlying scene graph. This scaffold will define the structure of the story and the form of the user interaction and will encapsulate whatever meta-scenario the developer wishes to deliver. The user will be presented with an interface with which to navigate through the scene graph, thereby defining a scenario, and adding a narrative to that scenario. Data capture will be used such that the result is a single defined scenario and a narrative dataset.

4.1 Functional requirements for developers

A requirement for the developers is an interface into which the non-linear plot nodes of the scene graph are defined. In the construction of each plot node the

developer of the scene graph can include an image, informative text and a text box for data entry. These items can be used individually or combined, depending on the needs of the scene graph being developed. The software must present these nodes to the user as a panel from a comic book, and include whichever elements the developer has chosen to include. Additionally, functionality is required that will allow the developer to link plot nodes to one another to provide the branching, non-linear basis of the story which will be formed by the user.

4.2 Functional requirements for users

The user of the application will build upon the scaffold provided by the developers. The first requirement is a method of presenting to the users the contents of the scene graph designed by the developers prior to the users interaction. An interface is required to display the content of each plot node as a cell, showing the images, text and an input area for the user to type in the narrative content. Secondly, once the user has completed one cell/plot node they must choose what happens next, constrained by options defined by the developers.

4.3 The Software

The application at its most basic level is a story node viewer; Figure 2 shows how the user interface is organized.

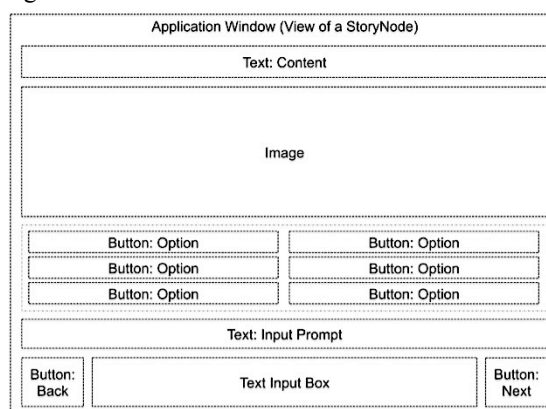


Figure 2: The application layout.

The scene graph, which is defined in an XML file and loaded at runtime, is essentially a list of story nodes. The following example code gives an example of a scene graph definition in which the user is presented with the beginning of the story "Jack and Jill" and asked to fill in a missing word.

```

<SceneGraph title="Jack and Jill"
identifier="001">
  <nodes>
    <StoryNode identifier="001">
      <target>002</target>
      <image>image.jpg</image>
      <content>1) Jack and Jill went up the
hill to fetch a ____ of water.</content>
      <text-input>What did Jack and Jill go
to fetch?</text-input>
      <nav-options />
    </StoryNode>
  </nodes>
</SceneGraph>

```

The possible navigation routes through the story nodes are defined using two methods. Firstly, each node has a target attribute that either points to another node or points to the 'end' (designating it a terminating node). Secondly, each node can have up to six navigation option child-nodes, each of which targets a node. The user interface converts these XML elements into interaction components. The 'target' attribute is presented as a next button and the navigation options are presented as a set of option buttons. Pressing next or one of the options performs a data capture of the current node, moves the view to the appropriate node and refreshes the view.

In the current incarnation of the software the content of a story node is very basic, containing text and a reference to an image representing the scene. As the image can contain whatever the developers would like to present and the text prompt for text input can be anything, the distinction between the aspects of the 'story' that the developers define and the aspects that the user create is flexible. For example the developers could decide to have the users input dialog and put no dialog in the images, or they could have lots of dialog in the images and have the user write narrative prompts. This flexibility, at this stage of development, is useful as it allows exploration of whatever story constructs that might be applicable for the audience with a complicated authoring process. Future versions of the application will include a more integrated authoring user interface. This will seek to avoid an overly complicated scenario definition convention as this would not only be difficult to develop but is likely to make user interaction more problematic.

3.4 Data Capture

Data capture is entirely abstracted from scenario structure. When the user moves to another node, the current node is taken to be complete and anything the user inputted is captured. A user data object is created and stored in an XML file. This file takes note of the user's identity and references the

scenario to which this data applies. By taking this data file and combining it with the scenario definition for that file the scenario the user created can be reconstructed.

While capturing the user inputs other aspects of the user interaction are recorded. The user data mentioned above only captures the final path the user takes through the scene graph, it doesn't capture, for instance, if the user backs up and follows another route. To solve this problem the application keeps a 'complete' history that captures user data for each node but does so for every node the user passes through every time they pass through it, creating an arbitrarily large list of user data objects. For analysis purposes some other aspects of the user's interaction are captured. The length of time the user spends on each node and the number of edits the user makes to the text input box are recorded. These are envisaged as a way to get some insight into whether some nodes get more attention than others.

3.5 Future Development

The current version of the application is a functional prototype. It delivers what was envisaged as its primary functionality: presentation and data gathering. As such, it could be used as a final application, however, its main purpose so far has been for pilot testing.

We hope to improve the usability and functionality of the application by making it a web-based server side application. This would make it platform independent, allowing a unified login system and centralizing the data gathering methods. This would also side-step the issue of access rights on user machines since data can be gathered by the server hosting the application rather than being 'saved' by the client machine.

4. CONCLUSIONS

In all our experiments children fully engaged with the process of completing or creating the narrative of a comic book, with pedagogically meaningful results. The processes that have lead to the creation of the application follow from what children have, sometimes unexpectedly, produced within these experiments. They have shown that it is possible to create an engaging activity that not only promotes literacy and literary skills in the creation of a narrative, but also allows for the development of concepts from other subject areas, in this specific

case inter-group sensitivity, in both individual and collaborative contexts.

In this paper we detailed the on-going development of an interactive comic book application for the scaffolding and creation of nonlinear stories. We have shown how the use of comic books and technology are beneficial to children's learning experiences. We also described two pilot studies in which we investigated an innovative approach to story creation through the use of comic book styled instruments.

ACKNOWLEDGEMENTS

This work was partially supported by European Community (EC) and is currently funded by the ECUTE project (ICT-5-4.2-257666). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

We would also like to thank the staff and pupils of St Mary's R.C. and Hudson Road primary schools, Sunderland, UK for taking part in the studies.

REFERENCES

- Abas, H. & Zaman, H.B., 2010. Digital Storytelling Design with Augmented Reality Technology for Remedial Students in Learning Bahasa Melayu. In Z. W. Abas, I. Jung, & J. Luca, eds. *Proceedings of Global Learn Asia Pacific 2010*. Penang, Malaysia: AACE, pp. 3558-3563.
- Birisci, S. & Metin, M., 2010. Pre-Service Elementary Teachers; Views on Concept Cartoons: A Sample from Turkey. *Middle-East Journal of Scientific Research*, 5(2), pp.91-97.
- Bitz, M., 2008. The comic book project. *School Arts*, 108(1), pp. 23-25
- Burnett, C. et al., 2006. Digital connections: transforming literacy in the primary school. *Cambridge Journal of Education*, 36(1), pp. 11-29.
- Cervesato, I., 2011. Discovering logic through comics. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education (ITiCSE '11)*. ACM.
- Cleaver, S., 2008. Comics & Graphic Novels. *Instructor*, 117(6), pp. 28-30.
- Figa, E., 2007. The Emergent Properties of Multimedia Applications for Storytelling Pedagogy in a Distance Education Online Learning Community. *Storytelling, Self, Society: An Interdisciplinary Journal of Storytelling Studies*, 3(1), pp. 50-72.
- Gerde, V.W. & Foster, R.S., 2007. X-Men Ethics: Using Comic Books to Teach Business Ethics. *Journal of Business Ethics*, 77(3), pp. 245-258.
- Herbst, P. et al., 2011. Using comics-based representations of teaching, and technology, to bring practice to teacher education courses. *ZDM*, 43(1), pp. 1-22.
- Jong, T., 2009. Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, 38(2), p. 105-134.
- Lenters, K., 2006. Resistance, Struggle, and the Adolescent Reader. *Journal of Adolescent Adult Literacy*, 50(2), pp. 136-146.
- Lopes, R. et al., 2009. Calligraphic Shortcuts for Comics Creation. In *Smart Graphics*. Springer, pp. 223-232.
- McTaggart, J., 2005. Using comics and graphic novels to encourage reluctant readers. *Reading Today*, 23(1), pp. 46-46.
- McVicker, C.J., 2007. Comic Strips as a Text Structure for Learning to Read. *The Reading Teacher*, 61(1), pp. 85-88.
- Moschovaki, E. & Meadows, S., 2005. Young Children's Spontaneous Participation during Classroom Book Reading: Differences According to Various Types of Books. *Early Childhood Research and Practice*, 7(1), pp. 1-17
- Norton, B., 2003. The Motivating Power of Comic Books: Insights from Archie Comic Readers. *The Reading Teacher*, 57(2), pp. 140-148.
- Pelton, L.F., Pelton, T. & Moore, K., 2007. Learning by communicating concepts through comics. In C. Crawford et al., eds. *Society for Information Technology and Teacher Education International Conference 2007*. AACE, pp. 1974-1981
- Pennington, R., Ault, M. & Schuster, J., 2011. Using Simultaneous Prompting and Computer-Assisted Instruction to Teach Story Writing to Students with Autism. *Assistive Technology Outcomes and Benefits Focused Issue: Assistive Technology and Writing*, 7(1), pp. 24-38.
- Di Raddo, P., 2006. Teaching Chemistry Lab Safety through Comics. *Journal of Chemical Education*, 83(4), pp. 571-573.
- Robin, B., 2008. Digital Storytelling: A Powerful Technology Tool for the 21st Century Classroom. *Theory Into Practice*, 47(3), pp. 220-228.
- Ruchter, M., Klar, B. & Geiger, W., 2010. Comparing the effects of mobile computers and traditional approaches in environmental education. *Computers & Education*, 54(4), pp. 1054-1067.
- Sabeti, S., 2011. The irony of "cool club": the place of comic book reading in schools. *Journal of Graphic Novels Comics*, pp.1-13.
- Sachs, G.T., Candlin, C.N. & Rose, K.R., 2003. Developing Cooperative Learning in the Efl/EsL Secondary Classroom. *RELC Journal*, 3(1), pp.338-369.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.