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.

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:

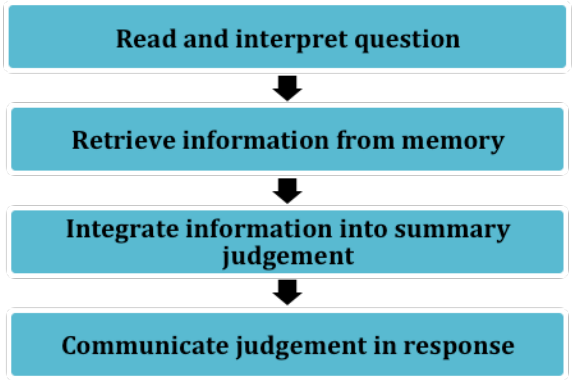

Figure 1: 4 stages of question answering [37]

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].

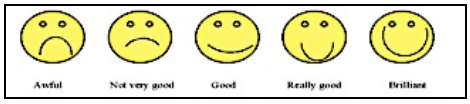

Figure 2: Smileyometer [29]

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).

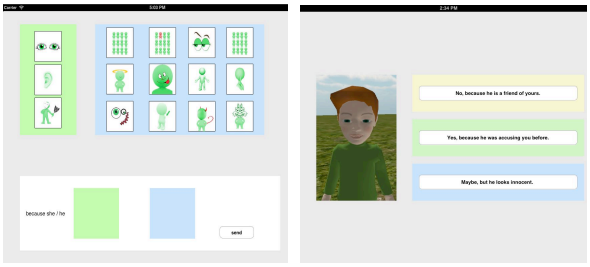

Figure 3: Screenshots of PIL-based interaction versus menu-based interaction

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.

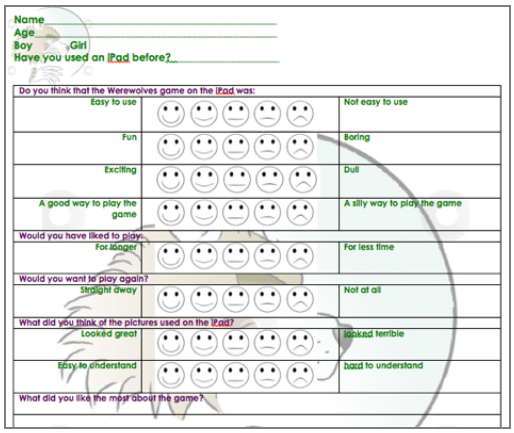

Figure 4: Pilot Questionnaire with traditional SFLs

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.

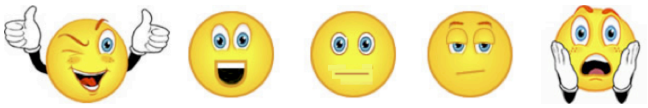

Figure 5: Dramatized SFL

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).

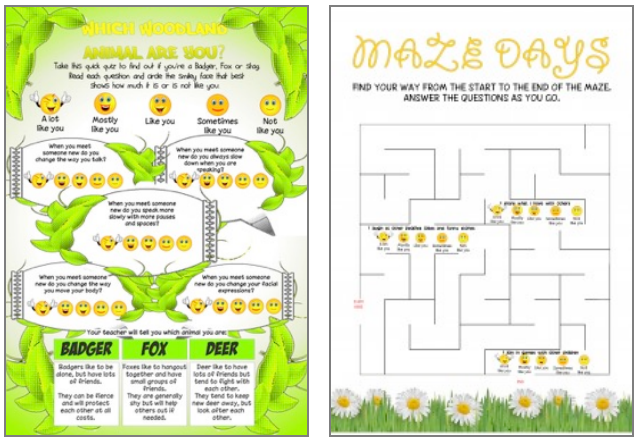

Figure 9: Pre- & post- test CQS

- 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).

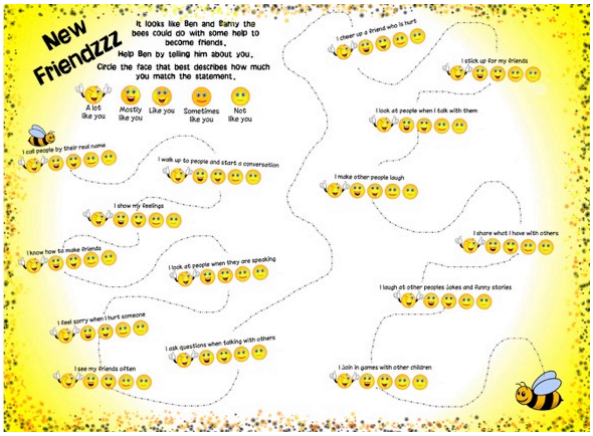

Figure 10: Pre-test MESSY

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.

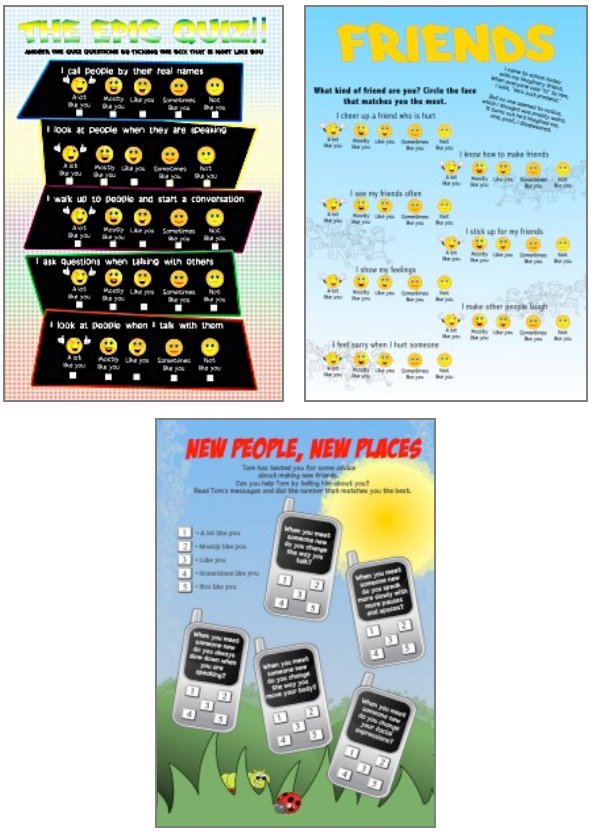

Figure 11: Post-test MESSY

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.

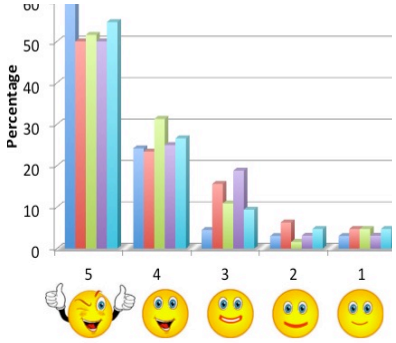

Figure 13: Children's views of the PIL

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 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 *Proceding 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 Mindfullness*. (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.