ABSTRACT
Over the past decade many new evaluation methods have emerged for evaluating user experience with children, but the results of these studies have tended to be reported in isolation of other techniques. This paper reports on a comparative analysis of 2 user experience evaluations methods with children. A within-subject design was adopted using 20 children aged between 7 and 8. The children played 2 different games on a tablet PCs and their experiences of each were captured using 2 evaluation methods which have been validated with children: the Fun Toolkit and This or That. The results showed that the Fun Toolkit and This or That method yielded similar results and were able to establish a preference for one game over the other. However, there were some inconsistencies between the results of individual tools within the Fun toolkit and some of the constructs being measured in the This or That method. Further research will try to identify any ordering effects within each method and redundancies within the questions.

Categories and Subject Descriptors
H.5.2 [User Interface]: Evaluation / Methodology

General Terms
Measurement, Design, Human Factors,

Keywords
User Experience, Evaluation Methods, Child Computer Interaction.

1. INTRODUCTION
Evaluation methods are well researched and understood in the HCI domain, and recently publications have sought to adapt existing methods to new domains [1] or improve the methodological procedures within the evaluation process [2]. This paper reports on a comparative analysis of two evaluation methods used with children to further enhance the body of knowledge within the domain. It is acknowledged that many of the traditional evaluation methods that have been deployed with adults are ineffective when adopted for evaluating technology with children. Hanna et al. outlined adaptations that would be required for standard usability testing procedures to work with children, including how the behaviour of the evaluator may affect the children’s performance [3]. Within the Child Computer Interaction domain a great deal of the early work focused upon the inclusion of children in the design process [4, 5] and as users in usability evaluation studies of technology [6-8]. Comparative studies emerged looking at the effectiveness of one method compared to another with children. For example the think-aloud protocol and post-task interviews were compared and the results showed no significant difference in the number of problems reported [9]. The emphasis over the last few years has moved away from usability evaluation methods focusing on the user experience. User experience has been described as the “new usability” [10] yet it has been suggested that user experience is not clearly defined or well understood within the HCI community [11] and it differs significantly in the constructs that are measured compared to traditional usability frameworks. ISO define user experience as a person’s perception and responses that result from the use and/or anticipated use of a product, system or service [12]. As a result new methods have emerged and the results of these have tended to be reported in isolation of other studies. Therefore it is important that comparative analysis is performed to understand the limitations, appropriateness in a given context in order to try and improve the methodologies.

Usability evaluations tend to focus on task performance whereas user experience focuses on lived experiences [13]. User experience is subjective and therefore cannot be captured by measuring task completion time or error rates which are traditional usability metrics. User experiences that can be captured could include physical, sensual, emotional and aesthetic experiences: for example if the objective of the evaluation were to measure fun then metrics would be required to capture these emotions. Carroll suggested that things are fun when they attract, capture, and hold our attention by provoking new or unusual emotions in contexts that typically arouse none [14]. Fun is one attribute of user experience that is important to measure as it is one of the major motivations for children to interact with technology [15], and Malone pioneered the study of fun as an important aspect of software [16]. Without the technology providing a positive experience children are unlikely to interact or accept it.

Several evaluation methods have emerged for use with children including Problem Identification Picture Cards [17], Fun Toolkit [18], Laddering [19] and This or That [20]. Whilst these methods...
have all been tested and validated with children this is done in isolation or against more traditional survey methods designed for use with adults. There have been no real comparisons specifically between the methods with researchers tending to choose just one bringing into question the reliability of an individual method and its results. If the different methods were to yield different results there is the potential for the results of entire research studies to be questioned and possibly be flawed. This could inturn be very costly both in time and money especially if the design of an application or piece of technology has been based on these results. It is imperative that the results of these different methods are similar, if not identical, when asking the same question using a different approach.

Many of these new methods rely upon the use of survey instruments or techniques and the use of survey methods with children often brings into question the validity and reliability of children responses [21] due to their cognitive and developmental abilities and there are large differences in these abilities between children of the same age [22]. This can lead to well known issues such as satisficing, suggestibility and understanding[23]. Maximizing the reliability of children’s responses is vital to ensure validity and integrity of results and any recommendations or decisions made from them.

This paper aims to compare the Fun Toolkit, a collection of tools created to evaluate the user experience of children before, during and after an evaluation study and ‘This or That’, a pairwise comparison method designed to be a cognitively less demanding evaluation method with children than other methods such as Laddering. The comparison will test the validity of these methods focusing on the reliability of the responses gathered. To evaluate these methods the children will play two games on a tablet PC using the stylus for interaction rather than the traditional mouse to allow a more natural interface, helping to eliminate the effects of poor motor skills and reduce conceptual issues associated with the device. Research into user experience and games has mainly focused on positive experiences such as fun or enjoyment and this will be the focus of the research presented in this paper. This paper provides an overview of the two methods, discussed in section 2, the results of the two are analysed independently in section 4 along with a comparative analysis, and finally in section 5 the discussion followed by the conclusion.

2. Evaluation Methods

There are numerous evaluation methods that could be adopted for measuring user experiences, however, it is important that the methods had been validated with children, and therefore the Fun Toolkit and This or That methods were selected. These methods are similar in their approaches and the results would yield quantitative data enabling comparisons to be made.

2.1 Fun ToolKit

The Fun Toolkit comprises of a number of techniques for eliciting information from the participants. The first tool is the Smileyometer, this is a visual analogue scale with the coding based upon a 5 point Likert Scale, with 1 relating to ‘Awful’ and 5 to ‘Brilliant’ (see Figure 1).

The Smileyometer is usually used before and after the children interact with the technology. The rationale in using it before is that it can measure their expectations, whilst using it afterwards it is assumed that the child is reporting experienced fun. The Smileyometer has been widely adopted and applied in research studies to measure satisfaction [24] and fun [25] as it is easy to complete and requires no writing on behalf of the children.

The Fun Sorter requires children to rate the technology or in this instance game to a number of different constructs. The children would rank the game based upon the different constructs selecting which was best and which was the worst. An example of a completed Fun Sorter used to compare the two games in this paper is shown in Figure 2, where G relates to the guessing game (Figure 5) and M the matching game (Figure 4).

The final tool is the Again Again table is a table that requires the children to pick ‘yes’, ‘maybe’ or ‘no’ for each activity they have experienced. In this study the children were asked ‘Would you like to play this game again?’ and they had to respond accordingly. An example of the completed Again Again table can be seen in Figure 3.
2.2 This or That
The This or That method is a pairwise comparison scale and it has been argued that it is the least cognitively demanding questioning style for children [20]. The method is relatively new first being cited in 2009 and enables the evaluator to ask a series of questions to establish the preference between two products on a number of constructs. For example the evaluator might ask “which was most fun, this or that?” A child would simply point to either the product or a pictorial representation of the product to indicate their preference. In the method described by Zaman 5 questions were used:

- Which game was most fun?
- Which of the games would you like to receive as a birthday present?
- Which game would you like to take home?
- Which would you like to play again?
- Which is a little bit stupid?

These 5 items were validated and proven to be internally consistent therefore were judged to be appropriate. However the authors were concerned about the possible ambiguity of the last question, the word stupid may be interpreted as a positive or negative construct.

3. Method
For this study, a within-subject design methodology was adopted in which the user experience of two games was evaluated using two different evaluation methods.

3.1 Participants
The participants in this study were 20 primary school children from a single class at a UK school, the children were from Year 3 (aged 7-8). They took part in this study during a day of activities held at the University of Central Lancashire. They played the games individually, and there was one researcher for every two children. The two researchers who took part in this study all had experience working with and conducting evaluations with children of these age groups. However one of the researchers had no experience with the This or That method, and therefore a brief training session was completed prior to the study.

3.2 Study Design
The study aimed to examine the reliability of two user experience evaluation methods with children. The methods require the children to rate their experiences based upon a number of constructs.

The order the methods were performed was dictated by the characteristics of the method itself. The Smileyometer would have to be performed first as it measures their experience prior to and after playing the game. As the This or That method is a pairwise comparison then this would have to be performed after the second game had been played. After the second game counter balancing techniques were used between the Fun Toolkit and This or That methods. In some cases the children would complete the This or That method before they completed the Fun Sorter and Again and Again table. Also the order the games were presented to the children was counterbalanced to mitigate any learning effect or possible bias to the results based upon playing a particular game first.

The Smileyometer was used as a method for determining the child’s overall experience of the game as it is claimed to measure feelings or experienced fun [26]. For the purpose of this study to enable comparison with the This or That method, a presumption was made that it would relate to fun. The other tools would enable a preference for a particular game to be established and this could be used to compare preference with the This or That technique.

The This or That method aimed to measure five constructs related to user experience. The data from this study would be compared to the results of the Fun Toolkit to ascertain the reliability of the children’s rating.

3.3 Apparatus
The two games were all played on a HP Pavilion tx200 tablet PC, which has a touch-screen, stylus and keyboard. The children were required to play both games using the stylus to interact with the PC. The two game games were both mini games within the purple palace game and were judged by the researchers to be age appropriate and suitable for the evaluation. By having 2 different styles of games it was anticipated that children would have a clear preference, which could be measured for consistency.

3.3.1 Matching Game
The first game was a matching game (see Figure 4). The children simply had to select a tile and a picture would be revealed, they would select another and if they matched the score would be incremented. If they did not match the tile would return to its original position and the children could select another two.

3.3.2 Guessing Game
The second game was a guessing game (see Figure 5) and required the children to select one of the items, hair, eyes, nose and mouth and they would have six attempts at guessing what the person looked like behind the curtain. Once the children had selected an item from each row they would select the tick and this would tell them how many they got right. If they did not match the tile would resort back to its original position and the children could select another two.
3.3.3 Data Capture Forms
For the Fun Toolkit the researcher gave the children a pen a a data capture form to complete for the Smileyometer, Again Again table and the Fun Sorter (see Figures 1-3). The researchers captured the responses to the questions from the This or That method by recording them on a separate sheet noting down the child who was responding so this could be matched with the data from the Fun Toolkit.

3.4 Procedure
Each child played both games and the two evaluation methods were used with each child.

Before the child played the first game a picture of the interface was shown to them and the child was asked to complete the first Smileyometer to measure their expectations of the game before playing. From the menu screen within the purple palace game the child was then asked to select either the guessing game or matching game depending on the order the child was playing the two games. Once the child had completed the first game they then completed the rest of the Fun Toolkit before they progressed onto playing the second game.

Before the child played the second game the Smileyometer was used again to measure their expectations. After they had played second game either the Fun Toolkit was administered followed by the This or That method or vice versa. For the This or That method the children were simply asked each of the five questions relating to the different constructs and were required to point to a pictorial representation of the game to indicate their answer. The researcher captured their responses on a data capture form.

Overall, the procedure for each child was as follows:

1. Pre-game Smileyometer for game 1
2. Play game 1
3. Post-game Smileyometer for game 1 and complete Fun Toolkit
4. Pre-game Smileyometer for game 2
5. Play game 2
6. Post-game Smileyometer for game 2
7. Complete either This or That followed by Fun Toolkit or vice versa

Approximately 5-10 minutes was allowed to play each game, but this was flexible to allow children to stop earlier if they were bored, or continue longer if they were engaged. The total length of the study for each child was 20 minutes. After completing the study, the children were thanked for their time and moved on to another activity, and another group took their place.

3.5 Analysis
The data sets from the two evaluation methods were each analysed in the following way.

3.5.1 Fun Toolkit
All children managed to complete the Smileyometers before and after they played the games. They were coded in an ordinal way 1-5, where for example 5 represented ‘brilliant’ and 1 ‘awful’.

The fun Sorter results were coded as 1 for the highest ranked game and 0 for the lowest ranked game for each of the two constructs being examined Fun and Easiest to Play.

The final instrument in the Fun Toolkit, the Again and Again table, resulted in a score for each game with yes being coded as 2, maybe 1 and no 0. Therefore the game with the highest value would be perceived to be their preferred choice.

A cumulative score was calculated for each of the games based on an aggregation of the Smileyometer after they had played it and the responses to the Fun Sorter and Again and Again table. The maximum score a game could achieve was 5 and the lowest 1 (based on them just selecting awful from the Smileyometer).

3.5.2 This or That
The This or That method is a pairwise comparison and was therefore coded in a dichotomous manner for each of the 5 constructs. When a preference was reported it was given a score of 1 for each of the first four constructs. The last construct stupid was judged to be a negative response. If a child selected one of the games, this was initially coded with a 1 with the other game being coded with a 0. These codes were then inverted to ensure that game which was not selected received a positive response.

Therefore the maximum score a game could receive was 5 which would indicate a strong preference for one game over the other.

4. Results
The results for the Fun Toolkit and This or That method are initially presented followed by a comparative analysis of the reliability of the methods.

4.1 Fun Toolkit
Each of the 20 children completed the Smileyometer before and after they played each of the two games and the results are presented in Table 1.

| Table 1: Mean Scores and Standard Deviation for Smileyometer |
|-----------------|-----------------|
|                 | Before | After |
|                 | Mean   | SD    | Mean   | SD    |
| Guess           | 3.95   | (0.89)| 3.55   | (1.15)|
| Match           | 3.80   | (1.11)| 4.40   | (0.68)|
For the Guessing game a t-test revealed a significant difference between the results of the Smileyometer before and after they had played the game $t(19)=3.629, p=.017$, suggesting their expectations had not been met. In contrast for the Matching game their initial expectations appear to have been surpassed as the t-test revealed a significant difference between the two results $t(19)=-3.559, p=0.002$. To determine whether there was a difference in user experience between the two games a t-test was performed, this was based upon the results of the Smileyometer after they had played the game and there was a significant difference between the two $t(19)=3.847, p=0.001$. It is evident from the Smileyometer that children had a clear preference for the Matching game.

The results for the Again and Again table are shown in table 2 below.

### Table 2: Frequency response to whether they would like to play it again

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>Maybe</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guess</td>
<td>5</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Match</td>
<td>16</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

It is clear from the results that the majority of children had a preference for the Matching game, with no child indicating that they would not like to play it again. The Fun Sorter looked at two constructs most fun and easiest to play and the results are shown in Table 3.

### Table 3: Number of children who selected a preference for a game based on the two constructs

<table>
<thead>
<tr>
<th></th>
<th>Guessing</th>
<th>Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Fun</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Easiest to Play</td>
<td>2</td>
<td>18</td>
</tr>
</tbody>
</table>

In line with the other results reported for the Fun Toolkit the Fun Sorter identified a preference for the Matching game for the two constructs.

The mean cumulative scores were calculated based on the combined results of the tools within the Fun Toolkit with a maximum value of 9 and a minimum of 1. For the Guessing game the mean cumulative score was 5.10 ($SD=1.1714$) whilst the matching game was 7.65 ($SD=.875$). A t-test was performed to compare the means of each game, there was a significant difference between the two $t(19)=6.265, p<.001$. The combined tools cumulative score within the Fun Toolkit clearly show that the matching game is preferred over the guessing game for the various constructs measured.

### 4.2 This or That

The results for each of the five constructs asked using the This or That method is displayed in Table 4. Two children did not provide a response to the last question so their data was removed for this construct and the t-tests below.

### Table 4: Number of children who selected a preference using the This or That technique

<table>
<thead>
<tr>
<th></th>
<th>Guessing</th>
<th>Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most fun</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Receive as a birthday present</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>Take home</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Play again</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Little bit stupid</td>
<td>17</td>
<td>1</td>
</tr>
</tbody>
</table>

For each of the 5 constructs it is evident that the Matching game was preferred by the children. A total score was calculated based upon the responses to each of the questions (with a maximum score of 5) and the mean score for the Guessing game was 1.33 ($SD=1.33$), whilst the mean score for the Matching game was 3.67 ($SD=1.33$). A t-test revealed that there was a significant difference between the two games $t(17)=-3.67, p=0.002$. From the This or That method it is clear that the children had a preference for the Matching game which was also evident in the Fun Toolkit.

### 4.3 Comparison between methods

Between the two methods there were a number of constructs that were identical for example the Fun sorter and the first question in the This or That method were trying to establish which game was most fun. Analysis was thus performed on the constructs that matched between the methods.

For the construct of fun, as analysed through the first question in the fun sorter and this or that method there was a consistency in response of 90%, in that 18 of the 20 children responses matched between the two tools.

In the This or That method the fourth question asked the children which they would like to play again this was compared with the results from the Again Again table. The Again Again table allowed the children to comment on each of the games thus allowing the children to state they would like to play both games again, whilst the This or That method was a forced response whereby the child had to state a preference for one game over the other. The results for this construct revealed a match in 13 cases out of 20 (65% match). In total 7 children stated that they would like to play one game again using the This or That method, though in the Again Again table they showed only stated maybe.

The final analysis looked at which game within each of the tools was preferred by each child, the results are presented in table 5, where D stands for a draw (indicating no preference was established between the two games), M was the matching game and G the guessing game.
To determine how the Fun Toolkit compared to the This or That method an overall preference was established by the frequency distribution between the tools. For example no clear preference could be established using the Fun Toolkit for Child 1 as there were two draws, whilst Child 2 the guessing game came out on top with it being reported the most fun in the Fun Sorter and the highest ranked in the Again and Again table. There were 14 children (70%) who reported a preference for the matching game over the guessing game. Despite this, there were a number of inconsistencies when looking at individual tools or constructs.

Table 5: Software which came out on top in each of the evaluation instrument.

<table>
<thead>
<tr>
<th></th>
<th>Smiley</th>
<th>FS1 Fun</th>
<th>FS2 Ease</th>
<th>AA</th>
<th>TT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child 1</td>
<td>D</td>
<td>G</td>
<td>M</td>
<td>D</td>
<td>G</td>
</tr>
<tr>
<td>Child 2</td>
<td>D</td>
<td>G</td>
<td>M</td>
<td>G</td>
<td>M</td>
</tr>
<tr>
<td>Child 3</td>
<td>D</td>
<td>G</td>
<td>M</td>
<td>M</td>
<td>G</td>
</tr>
<tr>
<td>Child 4</td>
<td>M</td>
<td>G</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 5</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>D</td>
<td>M</td>
</tr>
<tr>
<td>Child 6</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 7</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 8</td>
<td>M</td>
<td>G</td>
<td>M</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>Child 9</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 10</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 11</td>
<td>M</td>
<td>G</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 12</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>D</td>
<td>M</td>
</tr>
<tr>
<td>Child 13</td>
<td>D</td>
<td>G</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 14</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 15</td>
<td>D</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 16</td>
<td>M</td>
<td>M</td>
<td>G</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 17</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 18</td>
<td>G</td>
<td>M</td>
<td>G</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 19</td>
<td>G</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Child 20</td>
<td>D</td>
<td>G</td>
<td>M</td>
<td>M</td>
<td>G</td>
</tr>
</tbody>
</table>

5. Discussion

When comparing evaluation methods such as the Fun tool Kit and This or That a direct comparison of the performance can be difficult because of the variability in the elicitation methods, procedures and how the data is analysed. However when analyzing the results in this study it is clear that both methods are capable of measuring fun, both clearly identified a preference for the matching game over the guessing game. Despite this, there were a number of inconsistencies when looking at individual tools or constructs.

There does appear to be a certain amount of satisficing within the This or That method as over half of the children didn’t have matching results for the first four questions. It was expected that the first 4 questions would match based on our interpretation of the method as it was unclear from the literature whether stupid was a negative response. Table 6 below shows the percentage of children that the first four questions and first five questions matched.

Table 6: Percentage of children whose responses matched in the This or That method

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Match %</th>
<th>Not Match %</th>
</tr>
</thead>
<tbody>
<tr>
<td>First four the same, last different</td>
<td>45%</td>
<td>55%</td>
</tr>
<tr>
<td>All five responses the same</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

This raises questions such as why a child would find the guessing game most fun and also want to play this game again, but the same child would rather take home the matching game and receive it as a present. The Fun Sorter is very similar to the This or That method and it has been reported that using multiple constructs in the Fun sorter can lead to satisficing [19] therefore it is feasible that this may have occurred. It may be advisable to limit the number of constructs within the This or That method or randomize the order they are presented to the children to mitigate against this problem. To further analyse this issue a Cronbach’s Alpha was calculated for the scale based on the presumption that the last question was a negative value and the result was .734 which is just slightly lower than the results published by Zaman of .797 in the original study using this approach [16]. Another Cronbach’s Alpha was calculated with the assumption that Stupid was a positive construct and the alpha value decreased to .189 which would indicate the scale was unreliable. The overall presumption should be that Stupid should be treated as a negative construct. Overall the This or That method would appear to be internally consistent and valid for measuring user experience despite the potential issue of satisficing. The scale was further analysed to determine whether the reliability would increase if any items were deleted and this did not prove to be the case, for example with the question Which would you like to play again? removed, the alpha value would decrease to .685. Although the researchers perceived satisfying to be occurring from a statistical perspective the constructs within the This or That method appear reliable.

Literature highlights the need to be very careful with the language used when surveying children [18]. The use of the word ‘stupid’ in question 5 of the This and That method raises concern as it appeared that some children did see this as a positive construct.
whereas others saw it as a negative. This could introduce researcher bias if a child is unsure about the question and needs to have it explained further.

The Smileyometer (after) was compared to the overall score in the This or That method and there was no correlation. When analyzing the data from the Smileyometer alone it clearly establishes a preference for one of the games over the other, however it has been reported that children often have a tendency to select the extremes of the scale Brilliant or Very Good. This over enthusiastic response may account for the fact that there was no correlation as in some instance based on the This or That method a child’s total for the guessing game was 0 and 5 for the matching game (which shows a strong preference), whilst the score for the Smileyometer was 3 and 4.

The This or That method only enable you to determine if there is a preference for one game over the other it does not offer the opportunity to establish the magnitude of this preference compared to the Fun Toolkit. For example the Again Again table, a number of children stated that they would play both games again suggesting that they were enjoying both, whilst they were forced to choose one over the other in the This or That method. It is quite feasible that the child did not have a distinct preference for one game over the other.

When looking at individual responses it is evident that some of the children like both games, resulting in a disparity between the two evaluation methods. The total scores for the 4 children who preferred the guessing game according to the This or That method are shown in table 7.

<table>
<thead>
<tr>
<th>Child</th>
<th>Fun Toolkit</th>
<th>This or That</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child 1</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Child 3</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Child 8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Child 20</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

Child 1 according to the Fun Toolkit did not have a preference for either game which is a feasible result. However the This or That method forces the child to make an overall preference, and therefore, in this instance the guessing game came out on top. The results for child 8 are rather unusual as they have a strong preference for the guessing game using the This or That but prefer the other game using the Fun Toolkit. The discrepancy seems to be with the Smileyometer for this child as the before score was 4 and after score was 3, they stated they would like to play it again and according to the Fun Sorter the guessing game was most fun.

It needs to be noted when comparing the two evaluation methods that the two use different scales and yield different data however, several of the tools in the Fun Toolkit, and all questions in the This or That method, force the children to choose one piece of software over the other. Whist this does force them to show their preference related to the question being asked it does not take into account that they may in fact have no preference at all. This may well account for some of the individual discrepancies within the data set. Dhar and Simonson [27] highlight this issue in survey studies carried out with adults and the potential of forced choice options to produce conflict in responses and also psychological discomfort to the participant [28, 29]. It is often the case that a person may not have actually have a preference between two items so why should they be made to choose one. This raises some ethical concerns surrounding the tools that researchers need to consider when designing studies.

When conducting research with children it is highly important that the wellbeing and happiness of the child is put first therefore the use of forced responses should be carefully considered if there is even a slim chance the question being asked could be causing some discomfort to the child. If we are asking a child which piece of software they would most like to take home, or which they find the easiest, should we not give them the option of saying both or neither rather than making them choose one when perhaps they do not wish to.

If forced response questions are being used it is essential to minimise the chance of conflict occurring within the answers given by each child and also ensure that if used they are producing the same answers when different methods are asking what is essentially the same question.

6. Conclusions
All methods clearly show a preference for the matching game, this would indicate that they are measuring the same construct.

In looking at overall preference there was a reliability of 70% across the tools. Further research is therefore needed to try and enhance the reliability of the tools through possible redesign or improved understanding of the constructs children understand.

It is clear that measuring specific constructs is problematic and care needs to be taken to ensure that they are not ambiguous. The construct *Stupid* appeared to be interpreted differently by the children and it is now presumed that this is a negative construct within the This or That method. The difference between the results clearly shows that there is a need to use more than one method to enhance the reliability of the results and any recommendations coming from them.

The Fun Toolkit offers an advantage over the This or That method in the fact that it allows children greater flexibility in their responses, they can actually have no preference. However the Smileyometer does appear to artificially raise the scores as children are often over enthusiastic in this measure. The data from the Smileyometer could be normalized to reduce this effect.

It is important that there is a coherent measure for fun in software for children, and in this study it is clear that both evaluation methods can establish a preference, but there are limitations. It is important that further research is conducted to enhance the methods to enable decision makers to understand attributes associated with fun in software for children instead of merely a preference for one game over the other.

7. Further Research
Further research will look at enhancing the This or That method by examining any possible ordering effect the questions have on the method. It will also examine the potential issue of satisficing and whether there is any redundancy in the questions. Satisficing could potentially be eliminated by modifying the procedure within the This or That method by incorporating distraction or conversation between the questions.
For the Fun Toolkit further research will look at any possible ordering effect, should the Fun Sorter be performed after the Again Again table, this is something that has not been examined to date. Further research should be conducted to try and establish a method to mitigate against the over enthusiastic responses to the Smileyometer, this may just be in the post analysis of the data or a refinement of the tool.

Finally research needs to be performed to establish a clear list of constructs that children understand and can be used within when evaluating software or technology with children of various ages. The Fun Sorter and This or That method rely upon asking a series of questions and these may be ambiguous leading to reliability issues in the data.

8. REFERENCES


[19] Zaman, B. and Abeele. V.V. 2010. Laddering with Young Children in User Experience Evaluations: Theoretical Groundings and a Practical Case. in IDC. Barcelona: ACM.


