http://dx.doi.org/10.14236/ewic/HCI2022.8

CHUX Toolkit: A Method to Evaluate Long Term User Experience with Children

Gavin Sim
University of Central Lancashire
Preston, UK
grsim@uclan.ac.uk

Matthew Horton
University of Central Lancashire
Preston, UK
mplhorton@uclan.ac.uk

This paper reports on a study to understand the effectiveness of using the CHUX Toolkit to evaluate long term user experience with children. A study was conducted over 5 weeks to evaluate Purple Mash using the CHUX Toolkit that comprises of a diary and graphing tool with interviews. The participants were 26 children, aged between 9 and 10 years, from a UK primary school. Three constructs were analysed: Enjoyment, Learning and Ease of Use. The results showed that children could independently complete the diaries and graphing tool to report their experiences of using Purple Mash. CHUX enabled the researchers to understand how their experiences changed over time and identify the reasons behind this. There was consistency between the data reported in the different tools. The contribution of this paper is a new method for evaluating long term user experience with children.

Children. Long-term user experience. Diary. Interviews. Retrospective.

1. INTRODUCTION

Research into user experience (UX), and the creation of UX methods, for and with children, is not a new field with child friendly survey (Barendregt, Bekker, & Baauw, 2008; Read, MacFarlane, & Casey, 2002), interview (Zaman & Abeele, 2010), and observational techniques (Sim, MacFarlane, & Horton, 2005) widely used by the Child Computer Interaction (CCI) community and beyond. However, little of this research has looked at the capture of long-term UX, instead tending to focus on the momentary experiences of children after the completion of a task or study (Horton et al., 2019). For example, Javora et al. (2019) examined the attractiveness of an educational game after 20 minutes of play whilst Sim et al. (2014) examined fun after 10 minutes of playing. These studies assume that the constructs under investigation are static, yet studies have shown the dynamic nature of user experience and preference over time. It may be necessary to understand whether these ratings remain consistent over prolonged periods of use. For example, a child may find a game fun for a short period, but it quickly becomes repetitive and this may not be understood through momentary evaluations.

Research tools available for capturing long-term user experience with children are diary methods and the MemoLine (Vissers, Bot, & Zaman, 2013). Diary methods are used to collect momentary experiences over a period of time which can then be used to identify trends, and changes in experiences during the prolonged use of a product or system. Whilst diaries can be an effective tool, they can be a time intensive method for children and their gatekeepers to commit to. In addition, diaries do not provide a retrospective view of experience from the point of view of the child. The MemoLine was developed to provide a low-cost method of evaluation over a longer period of time. The tool only needing to be used at the end of a study, but accuracy issues have been reported (Horton et al., 2019).

In this paper we present CHUX, a graphing based tool to measure long-term UX with children. CHUX is designed to provide a tool more closely aligned to adult based graphing methods such as the UX Curve (Kujala et al., 2011).

2. RELATED WORK

2.1 LONG-TERM USER EXPERIENCE

An analysis of the literature identified three dominant approaches to evaluating changes in user experiences over time: 1) Cross-sectional research designs; 2) longitudinal research designs using pre-post test, or more data gathering moments through repeated measures; and 3) retrospective recall (von

Proceedings of the 35th British HCI and Doctoral Consortium 2022, UK

Wilamowitz-Moellendorff, Hassenzahl, & Platz, 2006). Repeated sampling and retrospective surveys are the most widely used tools within the HCI community for capturing long-term UX (Kujala et al., 2011). Whilst it may appear on the surface that these methods are capturing similar data, this may not be the case.

2.1.1 Evaluating Momentary Experiences

Repeated sampling tools are used to collect momentary experiences over time which means they provide data at the time the participant has interacted with a product. This could be through the repeated administration of a survey, such as was used over a 3 month period to evaluate a m-health mobile app (Biduski et al., 2020). Gathering this data over a period of time allows the identification of changes and trends, but these are still limited to a set of momentary experiences. A retrospective tool makes the participant reflect on the whole experience of using a project after a specific time period. It is therefore likely that specific momentary highlights or problems may be missed using a retrospective method, but on the opposite side the final opinions of the user are likely to be more genuine to their long-term opinion of the product (Horton et al., 2019). Therefore, as an example, a participant may have identified a lot of issues in the moment, but in the long term their experience of using the product was generally positive.

The important question at this point is what is the purpose of the data being collected? If the focus is on UX problems then perhaps a momentary method may be more appropriate, whereas if the overall opinion of the participant is more important than a retrospective method would suffice. There is of course the big question of the money and time available to the research team and whether the trade-off of a less intensive retrospective method will provide the required data.

Methods of collecting momentary experiences over time tend to differ based on the specific period of time and the type of data being collected. For rich qualitative data methods such as the Critical Incident Technique (Flanagan, 1954), Experience Sampling (Hektner, Schmidt, & Csikszentmihalyi, 2007) and the Event Reconstruction Method (Karapanos et al., 2009) are all useful. For quantitative data methods such as the AttrakDiff questionnaire (Walsh et al., 2014) and the Fun Toolkit (Read, 2008) can be used.

2.1.2 Retrospective Evaluation Methods

In contrast popular retrospective measures include the UX Curve (Kujala et al., 2011), Draw UX (Varsaluoma & Sahar, 2014) and iScale (Karapanos, Martens, & Hassenzahl, 2012). These tools are similar in that they ask participants to sketch out their experience over time, differing in the amount of qualitative data they also collect.

2.1.3 Evaluation with Children

Notwithstanding the array of research methods on evaluating long-term UX with adults, it is still largely ignored in the subfield of Child Computer Interaction. There have been a number of studies that have used repeated measures to capture change in experience over time. For example, Barendregt et al. (2006) have used the Smileyometer instrument, which is part of the Fun Toolkit (Read, 2008), in a pre-post research design over a 3 week period. Angeliki et al. (2022) gave children a survey to evaluate their motivation for playing an educational game at the end of a 2 month period. These approaches, although both using survey tools, are requiring the children to make a judgement retrospectively at the end of the period, which maybe reliant on the child's ability to recall or they maybe rating just on their last experience.

For designers and developers, the reliance on a child's memory is problematic as this may be influenced by the peak-end effect. This theory highlights that a retrospective evaluation of an experience is strongly influenced by the peak, and end moment, of that experience. When applied to games the positioning of the challenge can influence the results (Gutwin et al., 2016), thus strong negative experiences towards the end could influence the overall rating. In addition, studies have reported on the problematic nature of children's memory, where children have claimed to remember experiencing events that they only thought about or that were suggested by others (Ceci et al., 1994). This brings into question the reliability of data gathered from retrospective methods with children. For children the reliability may not necessarily matter as their memories have the power to guide their future behaviour (Norman, 2009) but for designers and developers accuracy may be important. If children are recalling negative experiences and products are being redesigned based on this inaccurate data, it may cost organisations time and money without improving the overall user experience of their products. Therefore, a combination of retrospective and momentary methods may be required.

Whilst diaries have been used in UX evaluation studies in the field of CCI (Read et al., 2018), it is sometimes the parents that complete the diary based on the children's activities that day. In other cases, it is the child completed the diary. For example, the evaluation of an oral health educational video game, (Aljafari et al., 2015) and when evaluating e-books with children and Colombo & Landoni (2014) used the diary over a two week period as part of a mixed method approach. Diaries can be used for understanding both momentary experiences and changes in experiences over time (Markopoulos et al., 2008). It is apparent that children can successfully complete a diary without the aid of adult intervention. To date in CCI only one

tool exists for evaluating long term UX using retrospective techniques the MemoLine (Vissers et al., 2013).

2.2 MEMOLINE

The MemoLine tool was developed for use on a project to evaluate educational video games for children aged between 9 and 11 years old (Vissers et al., 2013). In this study the MemoLine was used with a small number of children (n=6) to evaluate a serious game over a 6 month period. It has subsequently been used to evaluate games within a school and home context (Sim et al., 2016). In both studies the results showed that the children were able to use the instrument and recall past events relating to the games involved.

The tool was designed based on adapting the UX Curve (Vissers et al., 2013) with the assumption that this instrument would need modifications to work with children. Three modifications occurred:

- Adjustment 1: One dimensional timeline
- Adjustment 2: Temporal recognition cues
- Adjustment 3: Game experience constructs

Adjustment 1 consisted of replacing the two-dimensional curve format with a one dimensional timeline. This timeline would represent the period of evaluation. However, because of this adjustment, an alternative was needed for the Y-axis that originally represented the type of experience (positive or negative) that occurred over time. The use of three colours (green, red and grey), was adapted with green representing positive, red negative and grey indicating a period of no use.

Adjustment 2, the MemoLine consisted of a one-dimensional timeline that was visually divided in weeks or months (according to the elapsed time). In order to facilitate the children's orientation in time, visual recognition points based on the child's personal activities that occurred within the evaluation period (like the child's birthday, holidays, or the start of the new school year) were added. This required the child to mark these events above the timeline (see Figure 1).

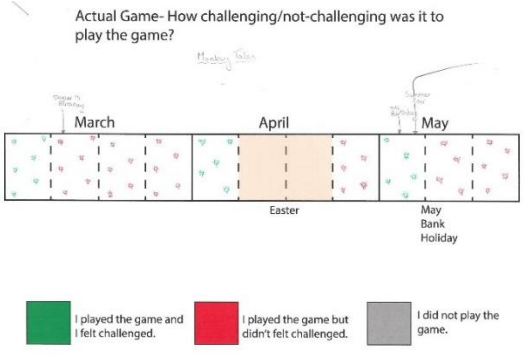

Figure 1: *Example of completed MemoLine (Sim et al., 2016)*

The final adjustment was the inclusion of game experience constructs with each construct having its own separate timeline. These constructs are valid if the evaluator is examining a game with children however for other technologies such as toys other constructs would be required.

In the studies that have used the MemoLine there has been concern about the accuracy of the data around periods of no play (Horton et al., 2019, Sim et al. 2016). For example, the child may report they did not play the game in week two for the construct fun and then indicate a positive experience in week two for challenge. This may impact the interviews, the validity of the findings, and important issues may not be identified. This highlights the further need for modification of the MemoLine, or alternatively that new methods are still required within CCI to retrospectively evaluate long term user experience.

3. CHUX TOOLKIT

The CHUX Toolkit consists of two components. The first is a diary that the children are expected to complete weekly. The aim of this is to act as a memory aid during the second part; a retrospective graphing exercise followed by interview questions.

3.1 Diary

The diary was designed to be simple and easy to complete by a child within a short period of time, in this case on a weekly basis whilst in school. The first page of the diary consisted of grid with one square representing a week (see Figure 2).

Figure 2: *Front page of the CHUX Diary*

This design is similar to how the MemoLine was constructed (Vissers et al., 2013). The decision was made to simply get the children to indicate whether they had a good or bad experience of using the product or system that week. The researchers were aware of the issues associated with using a forced

response as the child may have had a neutral experience that week and this cannot be reflected. The issues of forced response have been identified in other research tools notably the This or That method (Sim & Horton, 2012; Zaman, Abeele, & De Grooff, 2013) and the MemoLine (Horton et al., 2019). It has also been noted that children tend to categories experience as either positive or negative and this has been reflected in the design of the Smileyometer that has no neutral point (Read, 2008). Therefore, the decision was taken to use only two categories (good and bad) in the diaries, with the graphing tool facilitating a wider range of experiences.

The internal pages of the diary consisted of a page for each week with the same three questions on. The questions were:

- What I did
- What was good
- What was bad

Underneath each question was a text box for the child to write down their answer. The 'what I did' question was important in this context, as Purple Mash is a suite of educational resources and the activities changed during the 5 week period. This question could be changed to reflect the technology being evaluated, for example, in a game this might refer to the level of the game the player was on.

3.2 Graphing and Interview

The graphing tool was designed based on a combination of the UX Curve (Kujala et al., 2011) and the Smileyometer (Read, 2008). Children have successfully used the Smileyometer to rate their experiences on a wide range of different technologies and this was to be used on the y-axis with the time in weeks on the x-axis (see figure 3). The children would then indicate their experience on the graph for each of the 5 weeks using a different colour for each construct under investigation for example a green pen for learning.

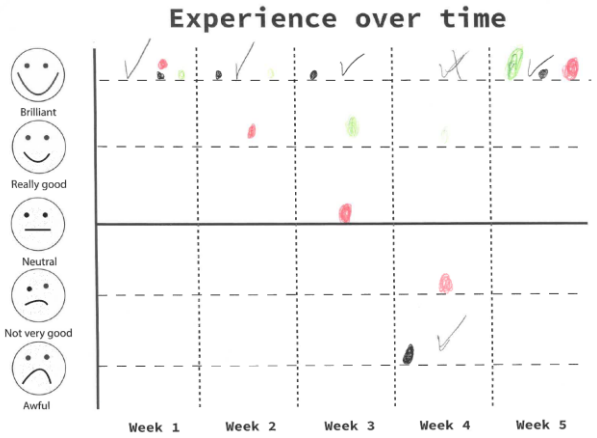

Figure 3: *Example of complete graph*

Using a similar interview style and questions to those from the MemoLine, questions are designed for each construct under investigation. For example, for enjoyment the questions were:

- How did you find the gameplay over the last 5 weeks?
    - How did the gameplay experience change over this time?
- What did you enjoy the most about the game play?
    - Could you tell me why?
- Which part of purple Mash did you enjoy the least?
    - Could you tell me why?

There were 3 questions for each construct and a follow on question to probe their answer further. Finally, there were 3 questions asked in order to evaluate the CHUX Toolkit to help the researchers understand any issues with either the diary or graphing component, the questions were:

- How did you find filling in the diary? What was hard or easy?
- How did you find the graph today? What was hard or easy?
- Did the diary or graph help you answer the questions? And why?

4. RESEARCH STUDY

This study aimed to investigate the effectiveness of the CHUX Toolkit for evaluating long-term user experience with children. The MemoLine tool was designed based on the assumption that graphing was not feasible for children and, as previously discussed, there were inconsistencies in reporting periods of no play within this tool. Therefore, the aim here was to explore children's ability to depict their experience over time within a graph, with an additional objective to understand whether children accurately reported periods of no play within CHUX.

4.1 Participants

The participants in this study were 26 primary school children from a UK school, the children were aged 9-10 years old. The total number of participants that completed the CHUX toolkit was 23, as 3 children were absent when the graphing and interview session occurred.

4.2 Apparatus

After consultation with the teachers in the school, the decision was made to use a website called Purple Mash (https://www.purplemash.com) which is a suite of games and applications designed to assist children in learning a range of subjects including

Maths and English. The software incorporated large tasks such as poster and blog creation with mini games that could be played to assist their learning of specific subjects. The children had access to the software within school and at home.

4.3 Procedure

The study was split into two phases the first was the introduction followed by graphing and interviews.

The introduction phase occurred one week before the start of the study. This approach has been taken in other studies (Horton et al., 2019), with the aim being for the children to become comfortable with the researchers before the actual study occurs. In this session children could play with some technologies such as brain interfaces and Raspberry Pi game emulators that they would not normally have access to. At the end of the session the researchers then explained to the children the research study they wished the children to participate in and showed them the paper-based diaries. This would enable the children to ask questions about the study to ensure that they can give ascent. The teacher was then responsible for ensuring that the children completed the diary on a weekly basis for the five weeks they were using Purple Mash.

In phase two the graphing and interviews took place six weeks after the initial visit. The graphing activity took place in the children's classroom using a whiteboard. Two researchers were present along with the teacher. The researchers introduced themselves again and the teacher gave the children their diaries to act as memory aids. One of the researchers explained to the children the aims of the study again, what research is within a university context, and sought to obtain ascent. Consent had already been obtained from the parents and guardian following ethical approval for the study from the universities ethics committee. The children were asked to review their diaries to aid them in recalling their experiences of using the software over the last five weeks. One of the researchers then used a PowerPoint presentation to demonstrate to the children how they might complete the graphing tool and provided opportunity for the children to ask any questions. The children were then given approximately 10 minutes to complete the graph for the three constructs. On the tables they were given different colour pencils to represent the different constructs and plot their experience on the graph:

- Enjoyment - red
- Learning - green
- Ease of use – black

Once this was complete the next stage was the interviews which were conducted throughout the day. The interviews took place in a corridor outside the classroom and children came in pairs to be interviewed individually by one of the researchers who noted their responses on a reporting form. When the children had completed the interview they were thanked by the researchers and they returned to class. Another pair of children would then be sent to be interviewed. This continued until all children in this group had been interviewed by the researchers.

4.4 Analysis

All the children managed to complete the diaries, but three children were not available on the day to complete the graph or be interviewed and their data was omitted. The final number of participants was 23.

When creating the graphing component of CHUX it was anticipated that this would be coded in the same way as the Smileyometer (Read, 2008) with score of 1-5 with each line representing the corresponding face. However, the children did not always mark on the line, marking both above and below it. Because of this, the decision was made to extend the coding to 1-11. The solid lines represented the even numbers whilst the spaces between the lines would be the odd numbers. For example, a mark in the box below the line next to the awful smiley was scored 1 and a mark on the actual line scored 2, whilst a mark in the box above the brilliant smiley was scored 11 (see figure 3).

5. RESULTS

5.1 Diary Results

All 23 children managed to complete the diaries over the five week period indicating a positive, negative or period of no play. Table 1 below show the number of children reporting a positive or negative experience on a week by week basis. The total number of responses each week should total 23 and as can be seen in week one the total is 22 meaning that one child did not use Purple Mash that week. Five children did not use purple mash for at least one week out of the five.

Table 1: *Positive (p) and Negative (n) responses from the diaries*

Week 1		**Week 2**		**Week 3**		**Week 4**		**Week 5**	
p	***n***	***p***	***n***	***p***	***n***	***p***	***n***	***p***	***n***
17	***5***	***15***	***6***	***17***	***6***	***7***	***15***	***18***	***3***

The number of positives experiences reported by the children ranged from 15-18 for four weeks whilst in the fourth week it was only 7. Only two children reported a positive experience on each of the five weeks and no child reported a negative experience for all weeks.

The diaries offered insights into why the children reported positive and negative experiences. For example, in week four, the children were interacting with something called *Jane's Monster* and this was a negative experience for 15 of the 22 children who were present that week. Reasons for this included:

- P16 - *"The 3rd question I was not doing anything"*
- P14 - *"You can't see you score"*
- P6 - *"It was way too hard on level 3"*

It was evident from looking at the responses that many of the children found the task to complex in week 4 which resulted in a negative experience.

5.1 Graph Results

All the 23 children managed to complete the graph using different coloured pencils to present the three constructs. This would suggest that children are capable of using the tool.

To understand whether children reported periods of no play consistently the diary was compared with the responses from the graph to look for consistency. Only one child (P1), who was absent for one week, then rated the Purple Mash on the graph for that period. The other four children who had a period of absence did not complete the graph for that period.

The children were asked to rate their experience for the three constructs over a five week period. On the graph the responses were scored between 1 and 11. Table 2 below shows the means and standard deviations for the construct enjoyment.

Table 2: *Means and Standard Deviations for the construct Enjoyment*

Week 1	Week 2	Week 3	Week 4	Week 5
7.39 (2.82)	***6.96 (3.21)***	***7.30 (2.49)***	***5.00 (3.58)***	***7.52 (3.26)***

A score of 6 represents the middle (neutral) value on the Smileyometer (see Figure 3) suggesting that, with the exception of week four, the children enjoyed Purple Mash.

A Friedman test revealed a significant difference between the ratings over the five week period $X^2(4)=10.392$, $p=0.034$. A Wilcoxon test was performed to determine where the significance was with the p values represented in table 3.

Table 3: *Friedman Test p values between weeks for the construct Enjoyment*

	Week 1	Week 2	Week 3	Week 4	Week 5
Week 1	***N/A***	***0.52***	***0.87***	***0.03***	***0.85***
Week 2	***0.52***	***N/A***	***0.859***	***0.09***	***0.52***
Week 3	***0.87***	***0.86***	***N/A***	***0.03***	***0.74***
Week 4	***0.03***	***0.09***	***0.03***	***N/A***	***0.01***
Week 5	***0.85***	***0.52***	***0.74***	***0.01***	***N/A***

A post hoc analysis was performed using a Wilcoxon signed-rank test with a Bonferroni correction applied, resulting in a significance level set at $p < 0.01$. There were only weeks four and five that demonstrated a significant difference $Z=-2.54$, $p=0.01$.

In the interviews children were asked in general about what they enjoyed with P10 stating they enjoyed the spelling test as they liked the sound effects of cheering when it was right. The spelling test was also popular with P16 whilst P2 also liked the graphics.

For the construct learning, table 4 below shows the means and standard deviations for each of the 5 weeks.

Table 4: *Means and Standard Deviations for the construct learning*

Week 1	Week 2	Week 3	Week 4	Week 5
7.39 (2.70)	***7.43 (3.22)***	***8.21 (2.11)***	***6.00 (3.13)***	***8.22 (3.00)***

Once again it was evident that in week four the children's scores were lower than the other weeks. From the interviews it was not always possible to match the learning activities to the week but in some instances the children provided this level of detail. For example, P19 stated that the activity in which you had to put the right words in the right boxes (nouns, verbs) helped them learn the most? Another child (P17) indicated that they learnt the most from the *Jane's Monster* game which taught them about adjectives. This was the game used by children in week four with and many reporting it glitching (including P17).

For the final construct, ease of use, the results are presented in table 5. Children appeared to struggle more with the software on week 4.

Table 5: *Means and Standard Deviations for the construct ease of use*

Week 1	Week 2	Week 3	Week 4	Week 5
5.96 (3.35)	***6.65 (3.39)***	***7.34 (3.11)***	***5.34 (3.65)***	***6.34 (4.03)***

In comparison to the other two constructs the scores for ease of use were lower except for week three. Once again week four scored the lowest for ease of use followed by week one. In the interviews following the completion of the graph, when the children were asked the questions relating to ease of use, they recalled a number of problems. For example, P8 stated that *"in weeks 3-4 the game kept crashing, the controls were hard to use and everyone told the teacher who had to help them"*. The game crashing was also reported by other children and this may have accounted for the low scores, for example P4 stated *"that it was not possible to type your answers"*. It maybe that the other two constructs influence their overall rating of Purple Mash when examining the results from the diary.

5.3 Children's reflection on CHUX

The final part of the interview related to asking the children about their experience of using the tools. Only 21 of the children completed this part of the interview. Each statement was analysed and coded as either easy, neutral or hard. Table 6 below shows the number of children who found completing the tools difficult, neutral or easy based on their responses to the interviews.

Table 6: *Number of children who found completing the tools easy, neutral or hard.*

Tool	Easy	Neutral	Hard
DIARY	***19***	***2***	***1***
GRAPHING	***19***	***0***	***3***

The majority of children found the diary component easy to complete. The child who stated it was hard struggled to think of things to write in the diary as they appeared to not have any issues. P15 stated *"Most weeks it was hard. Couldn't find problems except for week 4"*. This was reflected in the diary and the overall rating of the game. Where children provided more detailed discussion of why it was easy their rational was writing about what they did. P21 stated *"it was easy as I had already experience of what I was writing"*, whilst P22 *"got to write what I did"*. The diary component appeared to be relatively easy for the children to complete on a weekly basis.

For the graphing tool most of children found it easy to complete with only 3 children finding it difficult. The reasons they stated were:

- P11 - *"Pretty hard you had to choose"*
- P14 - *"Quite hard had to remember what you did"*
- P23 – *"You had to use your brain really hard to think"*

A few of the children appeared to struggle to recall events despite having access to their diaries. There were several children who reported completing the tools to be easy, however, very few children elaborating on why. One child, P2, stated *"Very easy, I did not get it at first but understood after explanation"*. Overall, it would appear that the majority of children were able to use the tools within CHUX.

6. DISCUSSION

This paper evaluated CHUX, a tool to evaluate the long-term user experience of children. There are several tools created for evaluating long-term user experience with adults and the only tool created specifically for child, the MemoLine (Vissers et al., 2013), assumed children could not use the graphing technique that these adult facing tools utilised. The CHUX tool consisted of a diary that would capture momentary experiences on a weekly basis, and a retrospective graphing tool.

The children were able to successfully complete the diaries with their experiences of using Purple Mash each week. Through the diaries it was possible to identify areas of the game that impacted their experiences such as the *Jane's Monster* game the children played in week four. When relying purely on the MemoLine (Sim et al., 2016) and retrospective tools the accuracy of the events is questionable. Out of the 23 children 19 reported the diary easy to complete with only one child having some difficulty. Their problems did not appear to relate to completing the diary but rather in what to write in it. If children are not experiencing difficulties, they may feel obliged to write something and this may cause anxiety. When asking for problems it may be useful to reassure children that they only need to report problems if they have experienced any.

The children only wrote one or two brief sentences in the diary (see Figure 4). For someone analysing the data it would be difficult to understand the problems with *Jane's Monster*, thus a mixed methodology is required to probe further. It is important to ensure that the diaries are kept brief to sustain competition rates, and in this instance, children were completing them in around 5 minutes in school. This was also important as it was not disrupting the school day and taking away time from other lessons.

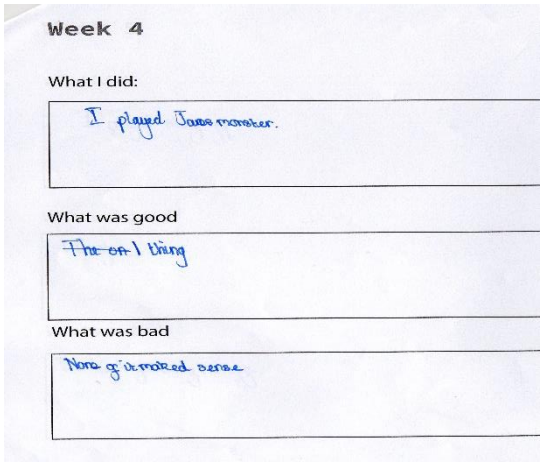

Figure 4: *Example of complete diary*

The diary was then used by the child as a memory aid helping them to reflect on their experiences prior to completing the graph and being interviewed. This graphing based tool designed to measure long-term UX with children aimed to be more closely aligned to adult based graphing methods such as the UX Curve (Kujala et al., 2011). All the children successfully completed the tool. In the study by (Horton et al., 2019) periods of no play were inaccurately recorded in on the MemoLine and this did not appear to be the case within the CHUX tool. Only 1 child inaccurately reported a period of no play using CHUX compared to 14 out of 22 using the MemoLine (Horton et al., 2019). This would suggest that CHUX offers more accuracy in the data collection process.

The graphing tool lets the children reflect on the entire period and with an 11 point scale offers children the opportunity to indicate the magnitude of their positive or negative experience, which is not feasible in the MemoLine. For example, child P21 weekly scores were 7,9,9,3,3 on the graph. This would show that the experiences were more positive in week two than week one, and very negative in weeks four and five. This subtle understanding of how their experiences may alter on a week by week is not feasible within the MemoLine where the child just reports whether it is positive or negative.

The interviews were a useful tool to probe further and understand the reasons their experiences changed over time that had not been captured within the other tools. For example, in week four, P21 stated in the diary nothing made sense, in the interviews further clarification was obtained and it was discovered that the answers were not being recorded correctly and as a consequence they could not complete two pieces of homework. Interviews have been shown to be an effective tool for understanding children's experiences (Zaman & Abeele, 2010) and this proved to be the case in this study.

Overall, the combination of tools in CHUX complemented each other to enable an understanding of how children's experience changed over time whilst using Purple Mash. Notably improvements in the accuracy of reporting of periods of no play were noted and greater understanding of the week by week differences could be achieved through the graphing tool.

6. CONCLUSION AND FURTHER RESEARCH

This paper presents a tool for evaluating long term user experience with children, CHUX. This is the second tool designed specifically for use with children and it has been shown to address many of the limitations of the MemoLine, notably the accuracy of reporting periods of no play. It has also demonstrated that graphing techniques are feasible for children to complete when evaluating change in experience over time.

CHUX consists of two components, a diary and a graphing tool that is followed by an interview, and the majority of children reported these easy to complete. The tool differs from the MemoLine in its addition of a diary component thus making it a hybrid evaluation methodology combining momentary and retrospective methods. However, care needs to be taken when designing the diary to ensure it can be completed quickly by the child and they do not feel anxiety if they can't think of anything to document. The combined data from the tool enable the researchers to understand how the children's experiences changed over time and the cause of this.

Further research will examine whether the graphing tool can be used in isolation of the diaries to make CHUX more versatile and aligned to adult methods such as the UX Curve. This study was carried out over a five week period and children may not be willing to complete a diary over several months, they may suffer fatigue. In addition, a prolonged evaluation period of months rather than weeks, may impact the graphing tool and other time periods are required along the x-axis to ensure children can successfully use the tool.

3 ACKNOWLEDGEMENTS

We would like to thank the children from class 6 and teachers from St Anne's Catholic Primary School for their participation in the study.

4. REFERENCES

Aljafari, A., Rice, C., Gallagher, J. E., & Hosey, M. T. (2015). An oral health education video game for high caries risk children: study protocol for a randomized controlled trial. *Trials, 16*(1), 1-10.

Angeliki, L., Rigou, M., Aliki, P., & Garofalakis, J. (2022). Effect of OSLM features and gamification motivators on motivation in DGBL: pupils' viewpoint. *Smart Learning Environments, 9*(1).

Barendregt, W., Bekker, M., & Baauw, E. (2008). Development and evaluation of the problem identification picture cards method. *Cognition, Technology and Work, 10*(2), 95-105.

Barendregt, W., Bekker, M., Bouwhuis, D. G., & Baauw, E. (2006). Identifying usability and fun problems in a computer game during first use and some practice:. *International Journal Human Computer Interaction, 64*, 830-846.

Biduski, D., Bellei, E. A., Rodriguez, J. P. M., Zaina, L. A. M., & De Marchi, A. C. B. (2020). Assessing long-term user experience on a mobile health application through an in-app embedded conversation-based questionnaire. *Computers in Human Behavior, 104*, 106169.

Ceci, S. J., Loftus, E. F., Leichtman, M. D., & Bruck, M. (1994). The possible role of source misattribution in th ecreation of fasle beliefs among preschoolers. *The international Journal of Clinical and Experimental Hypnosis, XLII*, 304-320.

Colombo, L., & Landoni, M. (2014). *A diary study of children's user experience with EBooks using flow theory as framework*. Paper presented at the Proceedings of the 2014 conference on Interaction design and children, Aarhus, Denmark.

Flanagan, J. C. (1954). The critical incident technique. *Psychological bulletin, 51*(4), 327-358.

Gutwin, C., Rooke, C., Cockburn, A., Mandryk, R. L., & Lafreniere, B. (2016). *Peak-End Effects on Player Experience in Casual Games*. Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, California, USA.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*: Sage.

Horton, M., Sim, G., Zaman, B., & Slegers, K. (2019). *Evaluating Long Term User Experience with Children: Comparing the MemoLine with Interviews*. Paper presented at the Proceedings of the 18th ACM International Conference on Interaction Design and Children, Boise, ID, USA.

Javora, O., Hannemann, T., Stárková, T., Volná, K., & Brom, C. (2019). Children like it more but don't learn more: Effects of esthetic visual design in educational games. *British Journal of Educational Technology, 50*(4), 1942-1960.

Karapanos, E., Martens, J.-B., & Hassenzahl, M. (2012). Reconstructing experiences with iScale. *International Journal of Human-Computer Studies, 70*(11), 849-865.

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J.-B. (2009). *User experience over time: an initial framework*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., & Sinnelä, A. (2011). UX Curve: A method for evaluating long-term user experience. *Interacting with Computers, 23*(5), 473-483.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., & Sinnelä, A. (2011). *Identifying hedonic factors in long-term user experience*. Paper presented at the Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces, Milano, Italy.

Markopoulos, P., Read, J. C., MacFarlane, S., & Hoysniemi, J. (2008). *Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers.* San Francisco: Morgan Kaufmann.

Norman, D. A. (2009). The way I see it: Memory is more important than actuality. *Interactions, 16*(2), 24-26.

Read, J. C. (2008). Validating the Fun Toolkit: an instrument for measuring children's opinion of technology. *Cognition, Technology and Work, 10*(2), 119-128.

Read, J. C., Horton, M., Clarke, S., Jones, R., Fitton, D., & Sim, G. (2018). *Designing for the 'at home' experience of parents and children with tablet games*. Paper presented at the Proceedings of the 17th ACM Conference on Interaction Design and Children, Trondheim, Norway.

Read, J. C., MacFarlane, S. J., & Casey, C. (2002). *Endurability, Engagement and Expectations: Measuring Children's Fun.* Paper presented at the Interaction Design and Children, Eindhoven, The Netherlands.

Sim, G., & Horton, M. (2012). *Investigating children's opinions of games: Fun Toolkit vs This or That*. Paper presented at the Interaction Design and Children, Bremen, Germany.

Sim, G., MacFarlane, S., & Horton, M. (2005). *Evaluating usability, fun and learning in educational software for children*. Paper

presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Montreal.

Sim, G., Nouwen, M., Vissers, J., Horton, M., Slegers, K., & Zaman, B. (2016). Using the Memoline to capture changes in user experience over time with children. *International Journal of Child-Computer Interaction*.

Sim, G., Read, J. C., Gregory, P., & Xu, D. (2014). From England to Uganda: Children Designing and Evaluating Serious Games. *Human–Computer Interaction, 30*(3-4), 263-293.

Varsaluoma, J., & Sahar, F. (2014). *Measuring retrospective user experience of non-powered hand tools: an exploratory remote study with UX curve.* Paper presented at the Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services.

Vissers, J., Bot, L. D., & Zaman, B. (2013). *MemoLine: evaluating long-term UX with children*. Paper presented at the Proceedings of the 12th International Conference on Interaction Design and Children, New York.

von Wilamowitz-Moellendorff, M., Hassenzahl, M., & Platz, A. (2006). Dynamics of user experience: How the perceived quality of mobile phones changes over time.

Walsh, T., Varsaluoma, J., Kujala, S., Nurkka, P., Petrie, H., & Power, C. (2014). *Axe UX: Exploring long-term user experience with iScale and AttrakDiff.* Paper presented at the Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services.

Zaman, B., & Abeele, V. V. (2010). *Laddering with Young Children in User Experence Evaluations: Theoretical Groundings and a Practical Case*. Paper presented at the IDC, Barcelona.

Zaman, B., Abeele, V. 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*(2), 61-70.