Small CCI – Exploring App Evaluation with Preschoolers

Janet C Read*
Child Computer Interaction Group,
University of Central Lancashire
jcread@uclan.ac.uk

Daniel Fitton
Child Computer Interaction Group,
University of Central Lancashire
dbfitton@uclan.ac.uk

Gavin Sim
Child Computer Interaction Group,
University of Central Lancashire
grsim@uclan.ac.uk

Matthew Horton
Child Computer Interaction Group,
University of Central Lancashire
mplhorton@uclan.ac.uk

Rhona Anne Dick
Lingokids
rhona.anne.dick@lingokids.com

Emanuela Mazzone
Lingokids
emanuela@lingokids.com

Rachel Forbes
Lingokids
rachel.forbes@lingokids.com

ABSTRACT

Child-Computer Interaction (CCI) is predominantly studied with school aged children. Working with preschool children, generally unable to read or write, involves addressing many challenges around planning, recruitment, and interpretation of findings. There are few examples in the literature of the challenges faced when conducting evaluations of technology with preschool children and very few evaluations conducted for commercial software companies. Our case study paper describes a six-week, twelve session, evaluation study of a commercial app (Lingokids) with children aged three and four in two nursery (preschool / kindergarten) schools. We describe challenges we met and describe how we adapted our plans to fit the context. We show how we were able to explore engagement and learning without gathering any personal data. With our practical tips and reflections, we hope our work will encourage others to work with young children in ways that respect their limited ability to understand assent and participation.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **HCI Design and Evaluation methods**; • **User studies**;

KEYWORDS

Preschool, Kindergarten, Evaluation, Ethics, Commercial, Case-study

ACM Reference Format:
Janet C Read, Daniel Fitton, Gavin Sim, Matthew Horton, Rhona Anne Dick, Emanuela Mazzone, and Rachel Forbes. 2023. Small CCI – Exploring App Evaluation with Preschoolers. In *Interaction Design and Children (IDC '23), June 19–23, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3585088.3589362

*Place the footnote text for the author (if applicable) here.

IDC '23, June 19–23, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0131-3/23/06.
https://doi.org/10.1145/3585088.3589362

1 INTRODUCTION

When Hanna and Risden [10] wrote their 1997 work on evaluating technology with children they probably didn't imagine user evaluations with children as young as three years. In 1997 preschool aged children were not using computer technology. Storey [26] reports how, in 1992, three schools in Newcastle, UK had installed BBC Microcomputers in nursery schools for a two year evaluation to determine their usefulness. When we look at computers currently used by young children the tablet/iPad format often proves most popular.

Since 1997 much has been written on evaluations of technology with children; new methods have been suggested (e.g. peer tutoring [11], Memoline [29], This or That [31]), and tutorials and courses have been delivered (e.g., [17], [25]), but most of this effort has focused on children between ages 5 and 11. The work on evaluating technology with young children is difficult to locate and often lacks practical tips to help new researchers navigate this exciting but challenging space. Very few papers report on the operation of commercially initiated evaluations where timescales, and focus, may be prescribed.

1.1 Young Children and Technology Use

The first challenge in exploring technology use with young children is to make sense of the definitions and terminology used in this space. We refer to young children as being those who are not yet in formal education and are typically unable to read. While limited by the physical and social infrastructures around them, and the level of funding / wealth available to their parents, these children typically spend some of their time in institutions that provide a structured environment for play and learning. This environment may be variously described as a nursery school, a kindergarten, and a preschool. When used in papers, these labels provide ambiguity, as, e.g. kindergarten in Finland ends when children reach 6, in the UK every 5-year-old is already in full time education. Research papers are sometimes rather unclear; Chang et al. [5] describe a study that at first sight is for kindergarten but uses children all aged

over 72 months, Beşevli et al. [3] were designing for children aged 3+ but were unable to access that age for their user study so used slightly older children. In the remainder of this paper, we refer to preschool and nursery interchangeably to refer to venues, in our case in the UK, where children aged 3 and 4 are learning through play.

1.2 Evaluating Technology with Young Children

Evaluation methods in CCI vary according to child-involvement, location, and design. There is an enthusiasm to develop commercially appropriate evaluation methods that parents, and teachers, can trust when selecting technology for the children in their care. One review describes 11 papers outlining rubrics and checklists for adults to use [22] including an 'app map' that guides an educator through an evaluation process [12], and the REVEAC tool that scores apps on educational content, design, functionality and technical characteristics [23].

In HCI the involvement of users in evaluation is considered core. With young children this involvement varies from them being observed using technology, to them being asked about technology and on to them engaging in user studies. An example of an observational study in a preschool is from Marco et al. [18], where technology was left in situ in a nursery so children could walk up and play with it as they liked. The Lanterns evaluation, in a Montessori nursery in the UK, was similar; giving children free exploration of the interactive technology [7].

In some studies, insights are gathered in a slightly more structured way; in one study of robots, 40 preschoolers, and a similar number of older children, spent 45 minutes with a robot before being asked questions about their experiences [15]. An interactive sleep companion was evaluated by comparing how children slept before they had the HugVie technology with how they slept after [21]. This method of evaluation was also seen in [28], who gave a pretest to 30 preschoolers who then met an IT intervention (an AR ABC app for an hour a day for a week) before doing a posttest. These sorts of studies, whilst showing 'significant' results, raise questions about the impact of the interactive technology on behavior change (e.g., the children may have just slept better because they were a couple of weeks older by then) and on learning (might the children not have learnt things anyway irrespective of the AR) as there is no 'control' group.

Control group studies have been reported; in one, children were divided into two groups with one group learning with an app for 45 hours and with the other group not seeing the app [8]. Using a pretest and posttest in this study showed an effect for the intervention. Similarly Haller et al. [9] worked with 42 children who each did one of three experiments. These designs raise some alarm bells about the ethics of such an approach as some of the children do not meet some or all the technology and so the principle of inclusion is not well applied. The only way such a study can be made inclusive is to rotate children so that each has a turn at each technology – this adds complexity and time to the study but is more ethical. A CCI challenge, especially for commercial work, is to have a study design that gives useful results whilst also ensuring ethical participation of children.

1.3 Challenges of Evaluating Technology with Young Children

As described above, studies do exist that show different ways that technology can be evaluated with young children, but in the main, there is very little written about the practicalities and challenges of working with this age of child. Challenges begin when contacting nurseries which may have regulations that can pose barriers for research work. In an evaluation of a reading app, researchers describe how the preschool they worked with only allowed children to be out of the classroom (on research activities) for 20 minutes at a time [27]. In another study, on HCI requirements for preschoolers, difficulties with getting consent forms were noted; despite giving out 24 consent forms, only 17 were returned [30]. Even once consented, drop out can be a problem; research on a new evaluation tool for young children was hampered by only getting ten children to stay the course of the study [19]. In a card sorting study, researchers had to access five nurseries to recruit enough children to give statistically reliable results [14]. There is a general feeling that much work on learning in preschool involves too few children and too many 'uncontrolled' studies [16]; robust studies are hard to manage with this age.

Whilst in the minority, papers do exist that critically evaluate evaluative work with preschool children. In a study, Antle et al., [2] highlighted the difficulties around trying to do control designs in field settings. In a case study around the evaluation of a software app [4], key pointers for conducting work with this age group were identified including bringing value, not just testing and running, designing research to bring minimal discord and maximal evidence, anticipating use scenarios and when measuring learning, being specific and acknowledging that children of this age cannot read. Our own work seeks to build on [4]'s observations on methods whilst also speaking to the concerns over the ethical inclusion of young children in robust evaluation studies.

2 CASE STUDY

We were asked by Lingokids to carry out an evaluation, with preschool children, of their existing app that uses playful learning to instruct children in early language skills. Lingokids wanted a controlled study, that complied with recommendations for children's screen time use and that obviously caused no distress to children. We were asked to compare the app (App) with two other modes of learning; learning using worksheets (Worksheets) and learning by being instructed by a 'teacher' using PowerPoint (PowerPoint) slides.

2.1 Study Design

The company and the researchers spent several weeks co-planning the study. Online meetings were held to discuss a protocol that would ensure the fair evaluation of the materials but also allow all participating children to experience all three conditions (App, Worksheet, PowerPoint) and all six learning topics *(ethics of inclusion), (robust evaluation).* The study would last 6 weeks with children attending twice a week, in twenty-minute sessions, with all the children working at the same time with university researchers. Thus, children would be typically in three groups of 4 – 5 children in three areas of the nursery school; on the first day of each week,

Figure 1: Set up for the True or False activity and a child completing a Worksheet on hand washing. Note the completed before and after Smiley faces also in the picture.

they would meet a new learning topic, on the second they would do reinforcement with no new material.

The initial plan was to break the children into 3 groups (A, B, C) with each group meeting 1 of the 3 conditions for 4 sessions in a row (viz. two weeks) before moving on to the next condition. Thus, group A might use the App for two weeks, the Worksheets for the next two weeks and then PowerPoint for the last two weeks, whilst group B would start with Worksheets, group C with PowerPoint and would then cycle same as group A.

As the company had familiarity with the content of the app and a robust understanding of the learning objectives, they developed instructional materials for the six learning topics which were recognizing feelings, counting, animals, body parts, colours and hand washing (Figure 1) supplying the research team with a comprehensive set of materials in the form of 12 worksheets and 12 slide decks (for introductory and reinforcement activities) and instructions on how to use them. They also delivering bespoke content to the app in such a way that children would just see the appropriate week's topic when they logged in.

To measure learning, we decided to triangulate measures of self-report of enjoyment from the children, using Smiley faces from the Fun Toolkit [24], teacher / researcher reports of engagement using the Leuven scale[1], and an assessment of 'momentary learning' after the second session of each week. In line with current thinking, a choice was made not to use a pretest/posttest model but rather to simply test learning in a playful way after the second session [20].

2.2 Ethics and Participation

The general principles around the ethics of doing HCI work with children demand that we pay close attention to 1. how children are treated, 2. how data is handled, and 3. how consent is gained. It is very important to treat children with care while they are doing a research study and close attention must be paid to their welfare – their own comfort must be put ahead of any scientific need for data. Personal data is governed by laws (e.g., GDPR) about its use and should only be collected when necessary. Whilst consent can be obtained from parents, the CCI community always seeks to gather assent from children – aiming to ensure they understand their participation is voluntary and their contributions are theirs to keep or give [1].

[1]https://learningjournals.co.uk/what-is-the-leuven-scale-and-how-to-use-it/

On the first of these points, we were happy that our design ensured that all children were well treated – they each interacted with all the conditions of the study thus ensuring inclusion, and we were content that our experience as a team (four researchers with over 50 years combined experience working with school children), and our belief in respecting children's contributions, would ensure a nice time for the children.

Personal data (2) and assent (3) went hand in hand in our thinking – the less data we gathered from the children, the easier it would be to inform the children about our intentions and thus, we could 'claim' a better version of assent; although in the action of doing a study we must respect any dissent from a child, even if inarticulate or poorly reasoned [13]. During the study we wanted to maximize opportunities for children to 'exercise agency', for example in deciding whether or not they wished to participate and in choosing the form of participation' as described in [6].

We decided to deal with data by collecting no personal data. Each venue we worked in collected in, and managed, all the consent forms and allocation of children to groups. We never knew any of the children's names (except when they announced them to us!) and we kept no records of which children were in which groups. In terms of data that we did collect, this was all captured on a group-by-group basis. Data collected would only be labelled with information about the session (date, location, group condition etc.) and nothing that could reveal the identity of a child was recorded. An example can be seen in Figure 1.

2.3 ACTIF - Momentary Testing of Learning

For each week of the study, we wanted to be able to capture a momentary learning score for each of the three conditions. We wanted this to be Anonymous, Comparative, Text-free, Inclusive, and Fun. Our aim was to gather a group score that was contributed to individually – we wanted to ensure that children were under no pressure and gave no personal data. The testing activities we designed, and their operation, are described in the next two sections.

2.3.1 Listen and Post. For five of the six weeks, we were testing the children's ability to match words and images; to test these we designed activities based on the posting of numbered cards into answer boxes (see Figure 2). We posed oral questions to children which they 'answered' by posting a numbered card in an answer box. The answer boxes represented items from the topics being taught (e.g., numbers, colours, animals etc.). Each numbered card related to a predefined question (and associated answer) that was read out to the children from a sheet. For example, the child might be asked to post their card in the box that represented the number 3. To discern differences between the conditions, we used different coloured card for each of the groups (A, B, C). Once all the children had participated in such a session, the research team emptied each box and made a count of the correctly, and incorrectly posted cards from each of the learning groups (see Figure 2 – right image).

In a variation to the use of boxes, one week we asked children to post stickers on body parts (see Figure 3) and used different coloured numbers to differentiate between the groups.

2.3.2 Right or Wrong. Whilst five of the learning activities (body parts, colours, numbers, animals, and feelings) lent themselves to

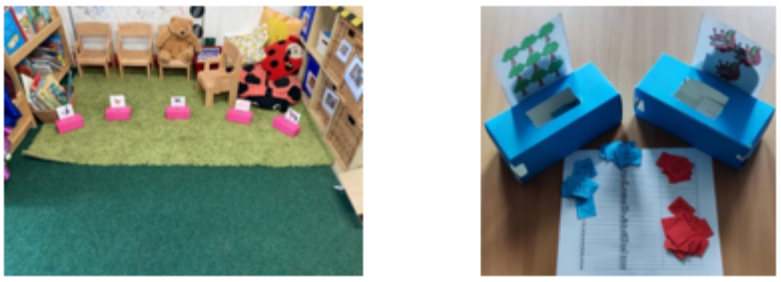

Figure 2: Boxes ready for children to post numbered cards into them and an example of a coding sheet used by researchers showing the collections of numbered cards from filled boxes.

Figure 3: Body parts – where 3 was arm, 8 arm, and 5 hand (all correct).

'*listen and post*', the topic one week was about hand washing. Here the children were required to learn the order in which to carry out activities (see the worksheet in Figure 1). To evaluate this activity, we asked the children questions about the activities *'what do we put on our hands before we rinse them?'*, then told them where to 'post' their answer based on whether their answer was correct or incorrect. For fun we used disposable cups which were posted into four coloured tubs; the facilitators knew which tubs were for correct and incorrect answers. We did not tell the children if their answers were correct or not, we simply said something like '*Great! put that in the '. . .green. . .' tub.*' In this way children all felt their answers were valued, there was no anxiety and the way we would count the number of 'correct' answers, was hidden from them. The different pictures on the cups let us know if the children's 'posts' were from the App, Worksheet or PowerPoint group (see Figure 1).

3 THE EXPERIENCE

3.1 Recruiting Children.

We recruited two nurseries to the study and immediately fell into our first difficulty which was the requirement to work with children for two sessions a week. When talking with the nursery managers it became clear that the attendance patterns of children would seriously limit the numbers of children available to us and would dictate which sessions would be most beneficial for us. As an example, one of the nursery managers painstakingly mapped out the children in nursery to provide a list showing how many children we would have if we did, for example, Tuesday afternoon and Thursday morning, or Monday morning and Wednesday afternoon. After establishing our timetables, we endeavored to attend the two nurseries on days that almost optimized our recruitment and the nursery managers then sent out consent forms to the children's parents.

3.2 First Sessions.

Because all the consent forms hadn't come in from the second nursery, we began week one with just one of the two nurseries. On arriving at the nursery for the first session, we found only eight children in attendance and realized that breaking these into three groups was not feasible. The research design needed quickly adapting so, there, and then, we decided to run two groups instead of three. The two groups were the App and PowerPoint learning. We made that decision knowing that we could compensate for this in the second nursery by running week one there with the App and Worksheet learning.

Two researchers worked with the two groups, whilst the nursery staff in the room engaged with the other children not participating in the study. Children first had to fill in a Smiley face to record how 'good' they thought the activity would be. It was clear several needed help with this, so this took longer than we expected, it then took a long time to log the individual children onto the App by which time the group doing PowerPoint had almost finished. Luckily there was a small library near the PowerPoint so the researcher with that group filled in time by reading to the children. Children filled in a second Smiley to indicate how much fun they had had, again with some not quite getting this, and the researcher with each group completed a score on the Leuven scale indicating each child's level of engagement. The children who had not played with the App, on that first session, immediately wanted to play with the App and were asking when their turn would be.

Two days later we had our first experience at testing learning which was of emotion faces and this went better than we expected. We laid out the six boxes with different faces on them representing different emotions, for example happy, sad etc. The two researchers brought their own groups of children to the room with the line of boxes, and each gave each child a single numbered paper (see Figure 2), one at a time, with the instruction to put that paper in the box labelled with a specific feeling such as 'happy' or 'sad' or 'angry' etc. Each number corresponded to a feeling on the master list that the researcher had (see example in Figure 3). Using boxes with slits ensured that children weren't looking inside or worrying about

what else was in the box. Each child got to post lots of numbers into the boxes, as there were 30 in total to post. Some children were less keen than others at participating and posted less; ethically they were able to 'opt out' of participation without any worry.

3.3 Adjustments without compromise

In our initial plan, some children would have had to wait four weeks before meeting the App. From our first session we realized that was far too long. The children learning with the App were visible to the other children as the nursery venues were open plan and we had to work in sight of the nursery staff to comply with child protection, so children were acutely aware of what they were missing, including the pink and blue earphones that the children got to wear when playing. We adjusted the protocol so that in each nursery one group would use the App while the other group had either the Worksheet or the PowerPoint activities and then the next week the groups would switch. This meant that all the children used the App every other week. We continued with the Smiley faces and did notice that as the weeks went on the children found completing them much easier and would complete them prior to prompting from facilitators; in other words - they had learned what to do.

3.4 Detail

Over the six weeks, two researchers were present at each session alongside nursery teachers. The four researchers attended both nurseries but in most of the sessions the same two researchers were in attendance. Over the 24 sessions carried out in total, the groups of children varied from as few as two to as many as seven with most sessions having four or five children in a group. We did not record individual ages of children, but all were aged 3 – 4. No video or audio recording took place. In each session we came into the nursery and said hi and then the children were organized by the staff, and we did the learning activities in two groups as described earlier – viz. complete Smiley (child), do instruction (researcher), complete new Smiley (child), complete Leuven scale (researcher). On the second session of each week, after the learning activities we did the ACTIF testing activity – for this we brought the children together, but each group worked under the 'direction' of the researcher who had been working with them earlier. On returning to the lab the results from the ACTIF testing, the Smiley scores, and the Leuven detail was analyzed by the researchers.

3.5 Findings

Some of the Smiley data had to be discarded as there were several instances when children had ticked all the faces, had coloured the sheets in or had given us something that we could not interpret. Overall, however, the reported fun from the children (N = 26) was highest when using the App (App: M = 4.3, PowerPoint: M = 3.61, Worksheets: M = 3.45). Engagement was highest with the App and Worksheets but not statistically different between the two. The learning data varied on a week-by-week basis and while there were some variations in the weeks, there was no significant difference found overall between the three modes of learning.

The Lingokids team were happy to know that children could learn from the App which was unsupervised and playful. Within the learning evaluation activities some data was lost from time to time; for example, on the week that involved the stickers, one child opted to stick the stickers on herself rather than the paper. It was evident that she knew the body parts, as when asked to stick it on the arm she would place it on her own arm, it was not possible to capture this data. The researcher opted not to force her to place the stickers on the paper.

4 KEY LEARNING

We developed a robust protocol that did not include gathering personal data (available at https://chici.org). Our methods of measuring learning were specifically designed with this in mind, and we encourage others to take this approach. Triangulating children's anonymous smiley data, anonymous scores on the Leuven scale and group scores for momentary learning gave numeric rigorous results for the company. In this paper we highlight how we had to plan and adapt with this age group. Despite piloting our materials with young children before we embarked on this study, we had not realized the possible upset that our design would cause with children working in open plan spaces. At both venues children were seen to be upset and not keen to engage if they did not get to use the App.

4.1 Practical Advice for Small CCI

In terms of practical advice, we offer the following:

- Be Very Prepared: Both before the activity and when setting up the activity, run through setup in advance, design in data collection safeguards, check/code the data soon as possible after the activity in case there are unanticipated problems that need to be resolved. Use debriefing sessions to reflect on the activities and be willing to adjust future sessions. In our work it often took several hours to design, pilot, prepare, print, and package materials for each 20-minute session.
- Be Playful: Make things into games, include novelty (i.e., posting things), include competition, explore new ideas. Children are naturally playful and if bored or disengaged can turn data collection into an unexpected game of their own invention (such as the example with the stickers), we may sometimes have to accept some data will not make it.
- Have Realistic Expectations: preschools are not homogeneous, be prepared to adapt data collection and accept that data collection might not be as rigorous as with older children. Participant numbers may fluctuate unexpectedly, and seemingly sensible data collection techniques may be reappropriated in unexpected ways (such as the child who placed stickers on herself instead of the intended sheet).
- Have a (Bold) Leader: Someone may need to change things in the moment. This could include the design of the study, for example reducing groups and adapting timings. Changes need to be rapidly made to maintain rigour in the study.

5 CONCLUSION

In a study evaluating a commercial app we were able to work with young children in preschool / nursery settings and gain data that could be used to draw defendable conclusions whilst ensuring that children could play, join in, not join in, express themselves and not be required to give any personal data (thus avoiding difficult

conversations about assent). Whilst we always asked children if they wanted to hand in the Smiley face grids that they had completed, we were happy for children to keep these and several did. We believe our ACTIF methods for measuring momentary learning, that ensured we could arrive at a score for a group whilst ensuring children did not feel under pressure to either participate or get a correct answer, were novel and were appropriate for a HCI style evaluation of learning. We do not suggest that these methods suit all contexts, they were bespoke designed for this evaluation. We have described some of the challenges that we encountered with the main aim to improve Small CCI that is HCI work with young children. In this case study we believe that we delivered 'minimal discord and maximal evidence' [4], whilst also delivering a playful engaging experience. The contribution of this work is the insights gathered from working with children and the reflections relating to the methodology.

Having worked with Lingokids on this project we feel we have learned skills in navigating a commercially initiated project and have also learned much about working with this age group and in preschool contexts. The work was complicated and required lots of preparation, redesign, and resilience but we believe that engaging with young children in the evaluation of apps is beneficial to the companies and the children.

SELECTION AND PARTICIPATION OF CHILDREN

To recruit children we wrote to, and visited nursery schools and preschools in the area and engaged with the first two who responded positively to our request. The managers of those establishments handled consent sending home to the children's parents our consent forms and information sheets as mandated by the University ethics approval which we had gained prior to the study. Children's assent was managed as described in the paper and no personal data was gathered. Children were free to drop out of participation if they so wished.

REFERENCES

[1] Alissa N Antle, Christopher Frauenberger, Monica Landoni, Jerry Alan Fails, Marina Jirotka, Helena Webb and Nalin Tutiyaphuengprasert. 2020. Emergent, situated and prospective ethics for child-computer interaction research. In *Proceedings of the 2020 ACM Interaction Design and Children Conference: Extended Abstracts*. (2020), 54-61.

[2] Alissa N Antle, Elgin-Skye McLaren, Holly Fiedler and Naomi Johnson. 2019. Evaluating the impact of a mobile neurofeedback app for young children at school and home. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. (2019), 1-13.

[3] Ceylan Beşevli, Elif Salman, Tilbe Goksun, Hakan Urey and Oğuzhan Özcan. 2019. MaR-T: Designing a projection-based mixed reality system for nonsymbolic math development of preschoolers: Guided by theories of cognition and learning. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. (2019), 280-292.

[4] Kristen Pilner Blair, Jay Pfaffman, Maria Cutumisu, Nicole Hallinen and Daniel Schwartz. 2015. Testing the effectiveness of iPad math game: lessons learned from running a multi-classroom Study. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. (2015), 727-734.

[5] Yu-Chen Chang, Jin-Ling Lo, Chao-Ju Huang, Nan-Yi Hsu, Hao-Hua Chu, Hsin-Yen Wang, Pei-Yu Chi and Ya-Lin Hsieh. 2008. Playful toothbrush: ubicomp technology for teaching tooth brushing to kindergarten children. In *Proceedings of the SIGCHI conference on human factors in computing systems*. (2008), 363-372.

[6] Sue Dockett and Bob Perry. 2011. Researching with young children: Seeking assent. *Child Indicators Research*, 4, 2 (2011), 231-247.

[7] Thomas Dylan, Abigail Durrant, Sena Çerçi, Shaun Lawson and John Vines 2021. Lanterns: Configuring a Digital Resource to Inspire Preschool Children's Free Play Outdoors. Association for Computing Machinery, Article 635. DOI:10.1145/3411764.3445745

[8] Saria Eltalhi, Huda Kutrani, Reem Imsallim and Mikal Elrfadi. 2021. The Usability of BenKids Mobile Learning App in Vocabulary Teaching for Preschool. *iJIM*, 15, 24 (2021), 5.

[9] Bettina Grünzweil and Michael Haller. 2009. Analyzing interaction techniques using mouse and keyboard for preschool children. In *Symposium of the Austrian HCI and Usability Engineering Group*. (2009), Springer, 448-456.

[10] L Hanna, K Risden and K Alexander, J. 1997. Guidelines for usability testing with children. *Interactions*, 1997, 5 (1997), 9-14.

[11] J Höysniemi, P Hämäläinen and L Turkki. 2003. Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game. *Interacting with Computers*, 15, 2 (2003), 203 - 225.

[12] Madeleine Heins Israelson. 2015. The app map: A tool for systematic evaluation of apps for early literacy learning. *The Reading Teacher*, 69, 3 (2015), 339-349.

[13] Steven Joffe. 2003. Rethink" affirmative agreement," but abandon" assent" (2003).

[14] Ana Vitoria Joly, Lyn Pemberton and Richard Griffiths. 2009. Card sorting activities with preschool children. *People and Computers XXIII Celebrating People and Technology* (2009), 204-213.

[15] Peter H Kahn Jr, Batya Friedman, Deanne R Perez-Granados and Nathan G Freier. 2004. Robotic pets in the lives of preschool children. In *CHI'04 extended abstracts on Human factors in computing systems*. (2004), 1449-1452.

[16] Junko Kanero, Vasfiye Geçkin, Cansu Oranç, Ezgi Mamus, Aylin C Küntay and Tilbe Göksun. 2018. Social robots for early language learning: Current evidence and future directions. *Child Development Perspectives*, 12, 3 (2018), 146-151.

[17] Stuart MacFarlane, Janet Read, Johanna Höysniemi and Panos Markopoulos. 2003. Half-day tutorial: evaluating interactive products for and with children. In *Interact*. (2003), 1027-1028.

[18] Javier Marco, Sandra Baldassarri, Eva Cerezo, Diana Yifan Xu and Janet C Read. 2010. LIFELONG INTERACTIONS Let the experts talk: an experience of tangible game design with children. *Interactions*, 17, 1 (2010), 58-61.

[19] Yusrita Mohd Yusoff, Ian Ruthven and Monica Landoni. 2013. Measuring emotion: A new evaluation tool for very young children (2013).

[20] Peter Moss. 2016. Early years PISA testing. *Early Years Educator*, 18, 6 (2016), 14-16.

[21] Junya Nakanishi, Hidenobu Sumioka and Hiroshi Ishiguro. 2016. Can Children Anthropomorphize Human-shaped Communication Media? A Pilot Study on Co-sleeping with a Huggable Communication Medium. In *Proceedings of the Fourth International Conference on Human Agent Interaction*. (2016), 103-106.

[22] Stamatios Papadakis. 2020. Tools for evaluating educational apps for young children: a systematic review of the literature. *Interactive Technology and Smart Education* (2020).

[23] Stamatios Papadakis, Julie Vaiopoulou, Michail Kalogiannakis and Dimitrios Stamovlasis. 2020. Developing and exploring an evaluation tool for educational apps (ETEA) targeting kindergarten children. *Sustainability*, 12, 10 (2020), 4201.

[24] J. C Read. 2007. Validating the Fun Toolkit: an instrument for measuring children's opinions of technology. *Cognition Technology and Work* (2007).

[25] Janet Read and Matthew Horton. 2022. Respecting and Facilitating Children's Contributions to Research, Design and Evaluation in HCI: Course Submission. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. (2022), 1-3.

[26] Gwyneth Storey. 1992. The educational value of computers in nursery schools. *Education 3-13*, 20, 1 (1992), 43-47.

[27] Anuj Tewari and John Canny. 2014. What did spot hide? a question-answering game for preschool children. In *Proceedings of the SIGCHI conference on human factors in computing systems*. (2014), 1807-1816.

[28] Neha Tuli and Archana Mantri. 2021. Evaluating usability of mobile-based augmented reality learning environments for early childhood. *International Journal of Human–Computer Interaction*, 37, 9 (2021), 815-827.

[29] Jorick Vissers, Lode De Bot and Bieke Zaman. 2013. MemoLine: evaluating long-term UX with children. In *Proceedings of the 12th International Conference on Interaction Design and Children*. (2013), ACM, 285-288.

[30] Monica Ward. 2014. HCI requirements for young primary school CALL learners. In *International Conference on Learning and Collaboration Technologies*. (2014), Springer, 339-348.

[31] B Zaman. 2009. Introducing a Pairwise Comparison Scale for UX Evaluations with Preschoolers. In *Interact 2009*. (2009), IFIP, 634 - 637.