View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CLoK

PLU-E: A Proposed Framework for Planning and Conducting Evaluation Studies with Children

Lorna McKnight
ChiCI Group
University of Central Lancashire
Preston, UK
lmcknight@uclan.ac.uk

Janet C. Read
ChiCI Group
University of Central Lancashire
Preston, UK
jcread@uclan.ac.uk

While many models exist to support the design process of a software development project, the evaluation process is far less well defined and this lack of definition often leads to poorly designed evaluations, or the use of the wrong evaluation method. Evaluations of products for children can be especially complex as they need to consider the different requirements and aims that such a product may have, and often use new or developing evaluation methods. This paper takes the view that evaluations should be planned from the start of a project in order to yield the best results, and proposes a framework to facilitate this. This framework is particularly intended to support the varied and often conflicting requirements of a product designed for children, as defined by the PLU model, but could be adapted for other user groups.

Keywords: Evaluation, framework, children, CCI, HCI, PLU model

1. INTRODUCTION

Evaluation is undoubtedly seen by the HCI community as a critical phase of product development, whether it be for assessing the suitability or success of a product, or to identify features in need of improvement or redesign. However, the process of evaluation is not clearly defined as it changes depending on the needs of the product, and as such it may be a difficult process to follow.

This paper arises from experiences encountered as part of the UMSIC[1] project. This is a trans-national 3-year EU-funded project that draws together researchers and developers from several countries with a wide range of skills and backgrounds. The aim of the project is to develop a mobile music-making application for children, with the intention of addressing social exclusion in marginalised individuals such as those with behavioural disorders such as ADHD (Attention Deficit Hyperactivity Disorder) or language difficulties. This means that experts from the fields of design, child psychology, music technology, usability and software development all need to work together to bring their expertise to the project. From a software development angle, integrating these individuals has proved difficult, as the different groups have different expectations and assumptions about the product development process. Developing a shared understanding is one of the great challenges of multidisciplinary research, and yet the benefits of bringing together influences from a range of fields can be immense, and so anything that can be done to make this process easier should be encouraged.

Part of the difficulty stems from a lack of formal models for the evaluation process. There exist many models of the design process (e.g. see Design Council, 2004), or of system development (e.g. see Sommerville, 2001), meaning there is some support for non-designers and non-developers to learn these processes. However, for designers and developers the process of evaluation may be less easy to understand. Evaluation is a word that means so many things to different disciplines, and so even the mention of 'evaluation methods' can cause confusion. In HCI, there is a wealth of information on evaluation methods, but far less information on how to choose the most appropriate method, and how this evaluation fits into the overall design process. While some project partners can be expected to have the expertise to choose the correct evaluation method, it can be difficult for them to explain to the rest of their interdisciplinary project team the rationale behind their choice, which leads to the common

[1]UMSIC Project: http://www.umsic.org

misconception that some methods are simply 'better' than others, rather than understanding the need to choose the best tool for the job. If the project partners do not have this expertise and therefore choose an unsuitable evaluation method, this can lead to weak evaluations that do not yield much useful data, and which are often performed at the end of development when it is too late to make improvements.

In the context of designing interactive products for children, this is even more critical, as evaluating for and with children is already a difficult process. This is discussed further in the following section.

2. EVALUATING WITH CHILDREN

There are many factors that can complicate the process of evaluating interactive products with children, and make it harder than evaluating with adults, meaning that studies often have to be planned with a bit more care.

As the field of Child-Computer Interaction (CCI) is relatively new as a field in its own right (see Read, 2005), there are few experts in child usability worldwide, meaning that most people working to evaluate software with children will be experts mainly in usability and user experience, or in education, or child psychology, or related disciplines. It can be hard for newcomers to the field to become aware of all the necessary issues they should consider. As for any distinct user-group, children have their own needs and requirements that designers and evaluators need to take into account, and for newcomers unused to working with children this familiarisation can be a daunting experience. The process of evaluating technology with children is discussed in more detail in Markopoulos, Read, MacFarlane & Höysniemi (2008), but some of the key issues relating to evaluation design can be identified as follows.

Adult evaluators will naturally have more access to adult participants than to children. Evaluating with children normally means recruiting known individuals, such as the researchers' own families, or working with a school, crèche or youth group. Fitting into a school's busy schedule can also be difficult, and combining the need to fit into a busy school day with the need to understand children's short attention spans often leads to studies being very short. Children also need some other motivation or reward for participating, such as making the studies fun for them, whereas for a school to become involved they may want to see a perceived benefit for the pupils, such as the tasks having an educational or skills-based value. They often also ask to let every child 'have a go' at the task, and so studies need to be short enough to allow time to include all pupils.

Apart from these practical difficulties, researchers are divided over the best methods for evaluating products for children, and many standard evaluation methods may not be appropriate. Many flaws have been identified with the use of survey methods (e.g. Horton & Read, 2008; Borgers & Hox, 2001), due to issues such as children misunderstanding questions, politeness, or simply a different understanding of the world. While some report success with verbalisation (or 'Think Aloud') with older children (Baauw & Markopoulos, 2004), it can be difficult for younger children (Donker & Reitsma, 2004). Observation methods are often used, but these require trained observers and always run the risk of imposing bias. These difficulties have led to more and more emerging methods being designed or adapted especially for children, such as the Fun Toolkit (Read, MacFarlane & Casey, 2002), the SEEM expert evaluation method (Baauw, Bekker & Barendregt, 2005), the This Or That pairwise comparison method (Zaman, 2009) and the Problem Identification Picture Cards method (Barendregt, Bekker & Baauw, 2008) to name but a few, but more work is needed to validate these methods and understand when they can and cannot be used.

This paper will not attempt to cover all possible evaluation methods that could be used. The important point to note is that a large number of methods exist, many of which have specific issues when used with children, and all of which have their advantages and disadvantages depending on the purpose of the study. As the field of CCI develops, new or altered evaluation methods also emerge frequently. Because of all this, it can be particularly hard to choose the right method to evaluate a product for children most effectively. In order to begin to do this, we first need to consider what the product aims to do.

3. CHILDREN'S VARIED ROLES: THE PLU MODEL

The PLU model (Read, 2004, in Markopoulos et al., 2008; discussed further in Read & Bekker, 2011) is a pre-existing model designed to assist in understanding and defining how children interact with technology. This model defines three different relationships that children have with interactive products, which map, in an approximate way, to the three genres of interactive technology. In this model, children are described as Players, Learners or Users, and the technologies are described as Entertainment, Education and Enabling. These represent three types of 'purpose' or position the

child may have when using technologies, which is expected to be mediated through a parent or teacher. The intended relationship of child to technology assists in considering how the interactive product might later be evaluated.

- Children as Players – in this relationship, the child should see the product as a plaything; to satisfy its purpose the product must **amuse** or **entertain** the child. Example technologies might include games or electronic pets.
- Children as Learners – the interactive product is seen as a substitute school or a teacher; it is expected to **instruct**, **challenge**, and **reward**. For example, this could include educational software or e-learning products.
- Children as Users – here, the child sees the interactive product as a tool; for the product to be useful it must enable the child and make things **easier to do**. Examples could include word processors, drawing tools or calendars.

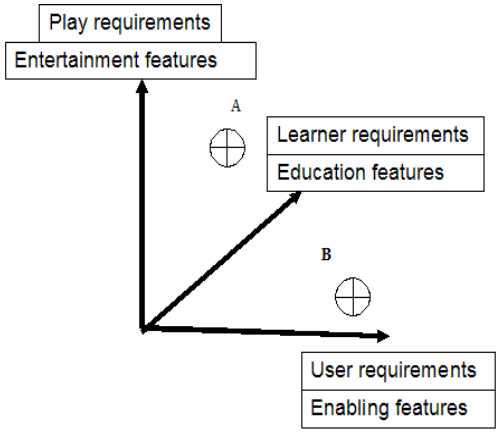

Figure 1: *The PLU model (Read, 2004, in Markopoulos et al., 2008)*

In the PLU model, a product and a child's purpose can be mapped in a three dimensional space, and the distance between these mappings can, to some extent, predict a mismatch between the designers' conceptions of the technology and the experience. This is shown in Figure 1 – if point A represents a child's perceived aims when using the system (e.g. to learn a little while having fun), then if the product being used by a child is positioned at point B (i.e. is designed to assist the child to carry out a functional task), the distance between the two is quite large, and thus it may not be a satisfying interaction.

Mapping a product onto this model begins to suggest how it might be evaluated. If the product is intended to primarily support entertainment requirements, it might be best to evaluate primarily in terms of user experience and fun. If the product is intended to support learning, it should be evaluated for pedagogical suitability and learning outcomes. If the product is mainly an enabling tool, it may be best to evaluate foremost for usability. This requires project teams to have a shared understanding of the aims of the developing product. It should be possible to give products a position on the 3D map, in terms of the levels of Playing, Learning and Using features that each product contains.

Following on from this, it is possible to start categorising types of evaluation methods, into those that are more suited to assessing the entertainment or experience aspects, those that are more suited towards assessing educational appropriateness, and those that are concerned with the usability of the product. Obviously there are several that fall into more than one category, but this may help to narrow down the choices that are available. A suggested categorisation of some popular evaluation methods in CCI is presented in Table 1, as an illustration of how this can work.

It is proposed therefore that the PLU model can be used as a key tool to assist in the choices of evaluation constructs. It is expected that most products will have features from more than one category, meaning that evaluators need to understand the intended weighting of these requirements, so they can tailor their evaluations accordingly. Based on these ideas, a design for an evaluation framework based on the PLU model can be suggested, which will be termed the PLU-E (Playing Learning and Using for Evaluation) framework. This will be outlined in the following section, before illustrating an example of how it could be used in a hypothetical project.

Table 1: *A sample categorisation of evaluation methods according to the PLU model*

Evaluating for Playing: considering fun, entertainment and experience	**Evaluating for Learning: considering pedagogy, effectiveness and learning outcomes**	**Evaluating for Using: considering usability, accessibility and efficiency**
Fun Toolkit (Read et al., 2002) This Or That (Zaman, 2009) Problem Identification Picture Cards (Barendregt et al., 2008) SEEM (Baauw et al., 2005)	HECE (Alsumait & Al-Osaimi, 2010) Pre and post tests (see for example Sim, MacFarlane & Read, 2006)	Think Aloud (see for example Baauw & Markopoulos, 2004) Problem Identification Picture Cards (Barendregt et al., 2008) SEEM (Baauw et al., 2005)

4. THE EVALUATION FRAMEWORK

The proposed PLU-E framework is as follows:

1. Decide on the purpose and/or focus of the product, both in terms of project goals and PLU. (For example is it meant to be a game with a little learning, a learning tool that is slightly fun, a supportive tool for specific children, etc.). This may already have been addressed to some extent in a requirements specification or project proposal.

    a) Are there parts of this (e.g. specific interaction techniques/interface components etc.) that present a particular challenge, and therefore need to be addressed as a priority?

2. Identify core users and specialist users.

    a) Are these discrete groups that should be treated separately, or all considered part of the user-group?

3. Based on stages 1 and 2, the project team agree on a PLU weighting that they feel represents the product. The team should agree as a group to what extent the proposed product aims to support Playing, Learning or Using. For example, an educational game for a standard group of children might be [P:30%, L:60%, U:10%] – the aims of the product are weighted higher towards learning but still require it to be fun; a homework diary for ADHD children might be [P:10%; L:20%; U:70%] – the product is an enabling tool, with a minor aim of teaching better practice; a game for pre-school children might be [P:60%; L:20% U:20%] and so on. This could be placed on the graph as a visualisation of this weighting.

4. Decide at what points in the process tests (evaluations) need to be made.

    a) It is expected that the tests will be 'feature-led', using throw-away prototypes or existing products. The focus and key features identified in stage 1 will be tested individually (e.g. if the product is a tangible game, one test may address the usability of the tangible interaction while another one may test the fun aspect of the game design, and so on), to lead into the end design of the product. It will also be specified how the data from these tests will be used to feed into the project.

    b) After testing product components, the project will then enter a prototyping phase, the length of which may vary depending on project constraints (e.g. this may involve paper prototypes, then a Flash version of the software, then a full implementation of key screens, and so on; alternatively it may be a full implementation from the start with some incomplete features that are added later).

    c) A 'final test' of the finished product should be planned at some point before the end of the project, giving time to fix problems (if the purpose of the project is the product development and release), or analyse the problems (if the purpose of the project is to research the process).

5. Based on stages 3 and 4 and project constraints (e.g. time and availability of users), the evaluations can be planned.

    a) For example a focus on usability may mean leaning towards evaluation methods that test this, e.g. inspection methods by experts (usability experts, developmental psychologists etc.) and user-testing. Evaluating fun is most likely done through user-testing with self-reporting and observations. Evaluating learning may be best done through expert evaluation by learning specialists (e.g. youth workers, teachers, educational psychologists etc).

    b) Each test in the process (as determined in stage 4) will need to revisit stages 1, 2 and 3, leading to different evaluations being chosen here. Each test will need to determine the most suitable form of evaluation for that component (e.g. in 4a the usability of a tangible game could be addressed through a heuristic evaluation or ergonomic testing of tangible artifacts, whereas the fun aspect might be tested through observations). It also needs to be asked as to whether it would be damaging for the users to view the product (or component) at each stage, or indeed damaging for the product (e.g. aspects that rely on novelty should not be introduced too early, or fun aspects should not be assessed if the product is unsafe).

    c) The final test (4c) should reflect the PLU weighting produced in stage 3: e.g. a product that is mostly intended to be fun should be mostly evaluated for how fun it is.

4.1 Example of the framework in use

This section gives a simplified example of how this framework could be used in a project, to form part of the project specification. The product described is a mobile music-making game – this is similar to the application being developed for the UMSIC project described earlier, but it should be noted that the details given here are hypothetical and are used purely for the sake of example.

1. The purpose and/or focus of the product.
The purpose of this product is a music-making application for children on a mobile device that aims for them to learn how to make music while

engaging with other users to increase social inclusion.
Key challenges: The product needs to address touch-screen interaction, social engagement and music-making.

2. Core users and specialist users
Core users are children aged 3-12. These can be divided into age-groups, who will not be expected to use the same product – these will be specified as a 3-5 group, a 6-9 group and a 10-12 group. Core users also include children with ADHD. These cannot be separated, as they are members of the same schools and so need to use the same product, but could be tested separately. Specialist users include teachers, who will have their own version of the product.

3. PLU weighting of the product
After discussion, the weighting was agreed by the project team as [P:60%; L:15%; U:25%] – the product is intended to be foremost a game, but needs also to be usable as a tool in music lessons. The learning goals of the game (i.e. music education) were not considered by the group to be as critical to the product as the entertainment aspects, which is why Playing is weighted higher than Learning.

4. Proposed tests and evaluations
Planned tests include 1) a test of music-making games [P:70%; L:20%; U:10%], 2) a test of touch-screen interaction [P:5%; L:5%; U:90%] and 3) a test of music teaching tools [P:5%; L:80%; U:15%]. After these, a prototype of the full product will be built [P:60%; L:15%; U:25%]. Two iterations of this prototype will be tested before the final version.

5. Plan of tests and evaluations
Test 1: music-making games – this will involve testing the fun level of different types of music-making games. Children in the 3 age-groups will each use a variety of similar games, then complete a self-reported assessment of the games using the Fun Toolkit, which was chosen as it tests the Play requirements of the game that this evaluation aimed to address. The data will help to determine the most desirable game features for each group.

Test 2: touch-screen interaction – this will involve expert evaluations of the device to be used, based on previous studies with children. This was chosen as the best way to test the Usable aspects of the game. The data will be used to inform the design of the interaction modes to be used.

Test 3: music teaching tools – this will include an expert assessment of teaching tools by music teachers. This will test the Learning aspect, and be used to discover the features most suitable for teaching music concepts.

Prototype test 1: a simple version of the product will be produced, and will be subject to expert evaluation. This will help to identify flaws to be fixed by the development team.

Prototype test 2: a version of the product will be produced on the intended device, and will be subject to user testing, using Wizard of Oz techniques for missing features. This will be used to test user acceptance and to identify simple changes that would improve the user experience.

Final test: the product will be user-tested as to how well it fits its proposed weighting [P:60%; L:15%; U:25%] – in other words it needs to be mainly fun, but children need to be able to make music, and should learn something from it. The evaluations for this will be weighted accordingly, and will be used to report on the success of the project.

5. DISCUSSION AND FUTURE WORK

This paper has aimed to show how evaluation can be made into a clearer and less daunting process for non-experts. By treating evaluations as a feature of the system that also needs to be carefully designed, it is hoped that this can lead to the most appropriate evaluations being carried out, and to avoid evaluations that yield little useful data.

The PLU-E framework described here does not attempt to dictate which method evaluators should use, recognising that they must use their expertise and domain knowledge in choosing the correct tool for the job, and that accepted practice will change over time. Instead, the framework merely attempts to guide the decision making process, in a manner that can be made clearer and more transparent for non-experts. This could also be assisted by developing a more comprehensive classification of evaluation methods, as illustrated in Table 1, and which are more suited to evaluating each of the three requirements of Playing, Learning or Using.

The framework as described here uses a percentage weighting system, to divide a total score between the three measures, rather than allowing each to be rated individually. This was based on tests where software designers were asked to rate software on how important each aspect was, and there was a tendency to see each aspect as equally important (for example rating each measure as 100%), which does not help guide the decision as to which is the most important, and therefore how to structure the evaluations. The percentage weighting system used here therefore is one attempt to solve this, but it may be that other rating mechanisms would be more effective for other project teams.

This proposed framework was designed by expert evaluators of children's technologies after conducting many evaluations on large-scale

projects, but it should still be noted that this framework is mostly untested at present. For it to gain acceptance, it would need testing on a real, large-scale project, so that the framework itself can be evaluated. Naturally this presents difficulties, as following a new methodology introduces risks to a project, and researchers may be reluctant to adopt untested methods. However, it should be remembered that this is not intended as a definitive description of how all evaluations should be performed, but an attempt to open the debate about the need for formal models of the evaluation process. By defining the process more clearly, it then becomes easier to change the process if necessary without causing undue confusion.

Finally, while this framework was developed specifically for the requirements of children, it is not inconceivable that the method could also prove useful when designing for other specialist groups. First, the needs of the user-group will need to be mapped out, in a similar way to the PLU model, after which the project team can agree on the focus of the product, and the evaluations can begin to be planned. The overall aim is that evaluations are always designed with the product's aims in mind, so that evaluators can gain the best possible data for their time. Hopefully this can then result in a smoother design process for the developers, and ultimately a better designed product for all users.

6. ACKNOWLEDGEMENTS

Part of this work has been performed in the 7th Framework Programme ICT project UMSIC (Grant Agreement No. 224561), which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in UMSIC, although the views expressed are those of the authors and do not necessarily represent the project.

With additional thanks also to all the ChiCI Group at UCLan, and all the staff and pupils at the schools that have worked with us.

7. REFERENCES

1. Alsumait, A. and Al-Osaimi, A. (2010) Usability heuristics evaluation for child e-learning applications. *Journal of Software*, 5 (6), pp.654-66.
2. Baauw, E., Bekker, M. M., and Barendregt, W. (2005) A Structured Expert Evaluation Method for the Evaluation of Children's Computer Games. In *Proceedings of Human-Computer Interaction – INTERACT 2005*, Rome, 14 September 2005, pp.457-469. Springer Verlag, Rome.
3. Baauw, E. & Markopoulos, P. (2004) A comparison of think-aloud and post-task interview for usability testing with children. In *Proceedings of IDC04*, Maryland, June 2004, pp.115-116. ACM, New York.
4. Barendregt, W., Bekker, M.M. and Baauw, E. (2008) Development and evaluation of the problem identification picture cards method. *Cognition Technology & Work* (10), pp.95-105.
5. Borgers, N. & Hox, J. (2001) Item non response in questionnaire research with children. *Journal of Official Statistics* 17 (2), pp.321-355.
6. Design Council (2007) *Eleven Lessons: managing design in eleven global companies.* http://www.designcouncil.org.uk/publications/Eleven-Lessons/ (retrieved May 2011).
7. Donker, A. & Reitsma, P. (2004) Usability testing with young children. In *Proceedings of IDC04*, Maryland, June 2004, pp.43-48. ACM, New York.
8. Horton, M. & Read, J.C. (2008) Interactive whiteboards in the living room? In *Proceedings of British HCI 2008*, Liverpool, September 2008, pp.147-148. British Computer Society, Swindon UK.
9. Markopoulos, P., Read, J.C., MacFarlane, S.J. and Höysniemi, J. (2008) *Evaluating Children's Interactive Products.* Morgan Kaufmann, Burlington MA.
10. Read, J.C. (2005) The ABC of CCI. *Interfaces*, 62, pp.8-9.
11. Read, J.C. & Bekker, M.M. (2011) The Nature of Child Computer Interaction. In *Proceedings of British HCI 2011*, Northumbria, July 2011. British Computer Society, Northumbria UK.
12. Read, J. C., MacFarlane, S.J. and Casey, C. (2002) Endurability, engagement and expectations: measuring children's fun. In *Proceedings of IDC02*, Eindhoven, The Netherlands, August 2002, pp.189-198. Shaker Publishing, Eindhoven.
13. Sim, G., MacFarlane, S. and Read, J.C. (2006). All work and no play: Measuring fun, usability, and learning in software for children. *Computers & Education*, 46 (3), pp.235-248.
14. Sommerville, I. (2001) *Software Engineering* (6th Edition), Chapter 3. Addison-Wesley, Essex.
15. Zaman, B. (2009). Introducing a Pairwise Comparison Scale for UX Evaluations with Preschoolers. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT '09)*, Uppsala, August 2009, pp. 634-637. Springer-Verlag, Berlin.