Comparison of Testing Environments with Children for Usability Problem Identification

Mohammadi Akheela Khanum [#1] , Munesh Chandra Trivedi [*2]
PAHER University
Udaipur, India
[1]akheela.khanum@gmail.com
[*] Dehradun Institute of Technology
Greater Noida, India
[2]Munesh.trivedi@gmail.com

Abstract— This paper report the results of an empirical study which was carried out to investigate the effect of testing environment on the results of usability evaluation process. The study involved 54 school children from India in the age range of 11-13 years. Children were asked to perform books searching tasks with International Children's Digital Library (ICDL). Children's activities with ICDL were captured by using CamStudio, an open source screen capturing software. The effect was quantified in terms of number of usability problems identified by the children when they were tested in various testing environments. The results are indicative that field testing with children can be a viable solution in terms of reduced time taken to complete the given tasks and reduced frustration levels reported by the children during the tests.

Keywords- usability evaluation; children; think-aloud; constructive interaction; problem identification;

I. INTRODUCTION

Children of today, born after the emergence of the Internet, are considered millennial [1] that have been born digital and raised as "Digital Natives" [2][3]. The design and evaluation of children's technologies have received increased attention during the last several years [4]. Children should be considered individuals with strong opinions, needs, likes, and dislikes, and they should be treated as such [5]. When evaluating technologies with children, evaluators are typically faced with unique challenges as children enter usability evaluation with special preconditions [6]. Thus, we need to understand how to create successful environments for children that facilitates usability problem identification.

Studies have shown that children are mostly affected by the context than adults. Context carries various meanings, in this study we refer to context as the physical location. The choice of location as context during usability evaluation is considered an important topic of discussion in research. Typically, the choice is between evaluating in an artificial setting such as a laboratory or in a more natural setting through a field evaluation.

Children show varying behavior when they are tested in the laboratory environment and when they are tested in the field environment. The importance of the physical context has been explored and studied by several usability researchers. However, we still lack clear empirical evidence of the merits of one environment over the other during usability evaluation with children. The purpose of our study is to find the answers to the following research questions:

RQ1: Are same usability problems found in both lab and field?
RQ2: Is the severity of the problems same in lab and field?
RQ3: Does the test environment affect the user performance?
RQ4: Which setting is more suitable to test with the children?

II. LITERATURE REVIEW

The importance of physical context in usability evaluation has been researched for a long. Out of the many factors that can effect usability evaluation, physical context is considered to directly influence the behaviour of the people involved in the usability evaluation. The physical context may include the location, the temperature, the time, the light etc. Some of the popular researches in this area are as follows.

Tullis et al. [7] compared remote and lab settings based on the time taken to complete the tasks and the problems discovered. Their study involved a prototype of a Web site for providing the employees of a company with access to information about their own benefits, including retirement savings information, pension information, medical and dental coverage, payroll deductions and direct deposit, and financial planning. The results found no significant difference between remote and traditional task times. Both remote and traditional lab testing revealed usability issues on existing websites. However, Tullis's participants scored the subjective tasks and interface differently between the different testing locations. It was concluded that the remote condition would incite

participants to be more honest regarding the test. Tullis offered no explanation for the difference in these scores, other than small sample size.

Tsiaousis & Giaglis [8] examined the effects of environmental distractions on mobile website usability. They hypothesize that environmental distractions can decrease user performance levels. They proposed a model hypothesizing on the effects of environmental distractions on the usability of mobile websites. They categorized the environmental distractions into auditory, visual and social. A preliminary test on 30 users was conducted to investigate the effect of environmental distractions on mobile website usability. Results confirmed that environmental distractions have direct effect on mobile website usability.

Hummel et al. [9] developed a mobile context-framework based on a small wireless sensor network, to monitor environmental conditions such as light, acceleration, sound, temperature, and humidity during the usability experiments. User experiments have been conducted in a laboratory with seven test persons where the environmental conditions were changed. Under varying environmental conditions the performance of the users on the average was decreased in terms of higher error rates and delays.

Andreasen [10] compared synchronous and asynchronous remote testing methods. Remote testing seemed to reveal interface issues. However, asynchronous study methods required more time to complete the tasks, and revealed fewer issues. However, asynchronous methods can be disseminated to larger groups, and the authors cite this as a benefit trade-off for asynchronous decreased performance. These findings were further supported by Bruun et al. [11], who found that remote, asynchronous testing identifies about half of the problems found by traditional usability testing, and their study concludes that the time savings introduced by the remote asynchronous method make them appealing for software usability testing.

Kaikkonen et al. [12] carried out usability testing of mobile consumer application in two environments: in a laboratory and in a field with a total of 40 test users. Results indicate that conducting a time-consuming field test may not be worthwhile when searching user interface flaws to improve user interaction. They found that field testing is worthwhile when combining usability tests with a field pilot or contextual study where user behaviour is investigated in a natural context.

Razak et al. [13] conducted usability testing with children in both laboratory and field. Drawing applications were tested in their preschool and an educational game was tested in the usability laboratory. The results indicate that field study is more suitable for understanding children experience with technology than it is with testing for usability problems and laboratory study is more suitable for evaluating user interfaces and interaction with the application than it is with understanding children's experience.

Andrrzejczak & Liu [14] conducted a study to evaluate the effect of testing location on usability test elements such as stress levels and user experience. A comparison between traditional lab testing and synchronous remote testing was conducted. The study investigated two groups of users in remote and traditional settings. Within each group participants completed two tasks, a simple task and a complex task. The dependent measures were task time taken, number of critical incidents reported, and user-reported anxiety score. Task times differed significantly between the physical location conditions; this difference was not meaningful for real world application, and likely introduced by overhead regarding synchronous remote testing methods. Critical incident reporting counts did not differ in any condition. No significant differences were found in user reported stress levels. Subjective assessments of the study and interface also did not differ significantly. Study findings suggest a similar user testing experience exists for remote and traditional laboratory usability testing.

Madathil [15] performed a synchronous remote usability test using a three-dimensional virtual world, and empirically compared it with WebEx, a web-based two-dimensional screen sharing and conferencing tool, and the traditional lab method. The study involved 36 participants in the test. The participants completed five tasks on an e-commerce website. The results suggest that virtual lab method is as effective as the traditional lab and WebEx based methods in terms of the time taken by the test participants to complete the tasks and the number of higher severity defects identified. Test participants and facilitators alike experienced lower overall workload in the traditional lab environment than in either of the remote testing environments.

Baillie & Schatz [16] evaluated a multimodal mobile application through a combination of laboratory and field studies. The users were given a set of four action scenarios to be performed. The results were surprising; only one action scenario was completed in the time frame whereas three out of four action scenarios were completed in lesser time. Error rates were higher in lab than in the field. The reason for such performances by the users could be that the users feel more relaxed in the field.

Karat [17] compared laboratory testing with the field testing. He applied the two approaches in order to help iteratively design a security application. Interestingly, participants completed the tasks in 25% less time in the field than when subjects completed similar tasks in laboratory conditions. Karat states that "there are possible

problems in comparing the results of the different tests; however the benefits of having both types of test data outweigh the negative factors ''.

Høegh et al., [18] investigated the role of field laboratories in evaluating the mobile systems. They evaluated several mobile systems in field settings over a period of four years. Findings of the study suggest that it is hard to evaluate mobile technologies in situ. It was difficult to capture the key moments of use and it was complicated to collect data of good quality. However, by means of a field laboratory with small wireless cameras and wireless microphones, it has been found that it is possible to capture field data about the use and usability of mobile technologies in a quality that matches that of a stationary usability laboratory.

Oztoprak & Erbug [19] compared laboratory and remote product usability testing. They tested a console type operator telephone set with a four row LCD screen. Five participants for each setting were involved in the test. The study found that usability testing and evaluation of a product in actual use context reveals implicit usability problems in the interface. Real tasks and real goals instead of simulated ones provide valuable usability information on product's usability on real use contexts. The post test questionnaire revealed that user prefers to participate in the test from their work places rather than travelling to the usability labs.

Kjeldskov et al., [20] compares the results produced by testing a mobile system in a laboratory setting and a field setting. Six participants in each of the two settings were assigned. They evaluated the number of usability problems found in the two different settings. Results reveal that conducting usability evaluation in the field has very little added advantage. Recreating central aspects of the use context in a laboratory setting enables the identification of the same number of usability problems.

III. CASE STUDY

The selected system for our experiment was ICDL. Fig. 1 shows the screen shot of ICDL website. This particular website was selected because digital libraries are becoming a common place for children and many researches are now focusing on how the children are using these new learning tools. During the children's demographic data collection, we also found that the children had never used ICDL.

ICDL is a collection of books that features various books for children in different age groups. ICDL has four search tools for accessing the current collection of books: Simple, Advanced, Location, and Keyword. Simple search allows the users to search for books using colorful buttons representing the most popular search categories. The advanced search allows users to search for books in a compact, text-link-based interface that contains the entire library category hierarchy. By selecting the location based search, users can search for books by spinning a globe to select a continent. Finally, with the keyword search, users search for books by typing in a keyword.

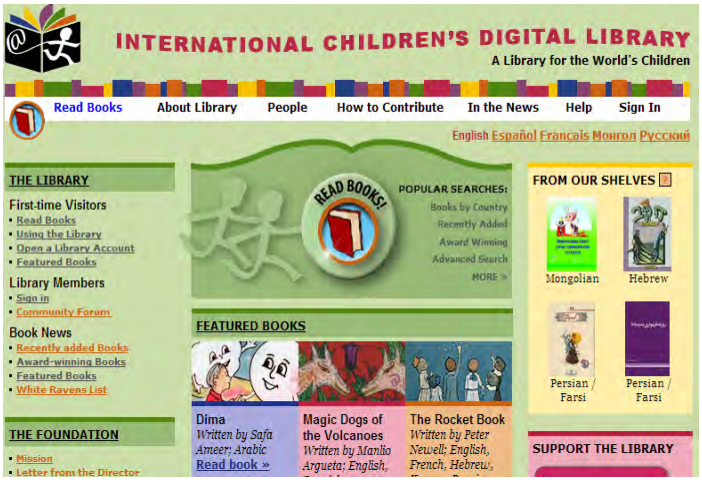

Fig. 1.Screen Shot of ICDL

A. Test Subjects

54 children (24 girls and 30 boys) at the age ranging from 10 years to 13 years old (M=11.63; SD=0.88) participated as test subjects in the experiment. All the children were 6th and 7th grade pupils from two different English medium schools in the Lucknow area of India. We did not compensate children for their involvement in the experiment. The children were assigned as test subjects to one of the four test setups: as individual testers in the lab and in the field for think-aloud sessions, as pairs in lab and field for constructive interaction sessions. Each individual setup had 9 individual testers (4 girls and 5 boys), and each paired setup had 9 pairs (4 pairs of girls and 5 pairs of boys). Children were randomly assigned to each of the four test setups. Children in pairs were familiar with each other. Table I shows the assignment of children to different setups.

TABLE I
54 Children Assigned as Individual Testers in Think-aloud and as Pairs in Constructive Interaction

	Constructive Interaction		**Think-aloud**	
	Lab	**Field**	**Lab**	**Field**
Boys	5x2	5x2	5	5
Girls	4x2	4x2	4	4
Total	9x2	9x2	9	9

B. Procedure

The sessions were held at the school's campus itself as we were denied permission from school authorities to commute to the place where the usability laboratory was located. Therefore, we created two labs in the school, one for field testing sessions, and one for laboratory testing sessions. For the field testing, we chose the school's computer lab with which the students were familiar and we tried to keep it as it was used by the children. No restrictions were imposed on the people to move in the lab during the test session. This created a perfect field environment for the children. For testing in lab environment, we setup a usability laboratory in one part of the school. The lab environment was kept different from the field environment. Lab was located in a quiet place where people not related with the test sessions were not allowed. The lab was only occupied by the test monitors and the test participants at any given time during the test sessions.

The first step towards starting the test was to take consent from the school authorities. After clearing the first step, we proceeded with taking the consent from the children's parents or guardians. To do so, we handed over the consent forms to the children to get it signed by their parents or guardians. The consent form provided information about the type of test their wards will be involved in and that the choice of allowing their children to take the test was purely voluntary. After receiving consent from 54 children, we scheduled the usability evaluation sessions. At the beginning of the test session children were introduced to the experiment by two of the participating researchers. The researchers explained the children's roles in the experiment and how their participation would contribute to our research.

Hanna et al. [6] guidelines for usability testing with children were followed. We greeted and children and introduced ourselves. Particularly, we focused on stressing the importance of the participation, and stressing that they were not the object of the test. The purpose of the usability test was explained to the children in detail. The children received questionnaires on which they had to provide answers to such as age, name, school, computer/internet experience, number of hours spend each week on computer/internet, and online reading experience. The usability test sessions were conducted in two labs, one a specialized usability laboratory setup in the school and the other was the school's computer lab. During the test sessions, all the screen activities and children's interaction with ICDL were recorded using CamStudio for later analyses. CamStudio is an open source desktop screen recorder.

The children were asked to solve five tasks. The tasks involved the use of different search options in ICDL. This included searching books by country, searching books by title, searching books by language, searching award winning books in English and reading a specified book in the language of their preference. We did not specify any time limits for the tasks, but required the participants to try to solve all tasks.

IV. DATA ANALYSIS AND RESULTS

All the raw video data was analyzed afterwards and a list of problems was constructed. The severity of each of the problems was categorized according to the definition by Rolf Molich [21]. According to the definition, a problem experienced by a participant falls in one of three categories:

• **Cosmetic:** The user is delayed for less than one minute, is mildly irritated, or is confronted with information, which to a lesser degree deviates from the expected.

• **Serious:** The user is delayed for several minutes, is somewhat irritated, or is confronted with information, which to some degree deviates from the expected.

• **Critical:** The users attempt to solve the task comes to a halt; the user is very irritated or is confronted with information which to a critical degree deviates from the expected.

The categorization was done by observing the video recording of each participant, and then evaluated each situation according to the guidelines described above.

The analysis of 36 usability test sessions resulted in the identification 121 different usability problems as depicted in Table II.

Table II
Number of Identified Usability Problems

	Constructive Interaction		Think -aloud		Combined
	Lab	**Field**	**Lab**	**Field**	
Cosmetic	11	10	11	11	43
Serious	16	17	10	15	58
Critical	4	4	3	9	20
Total	31	31	30	29	121

Our experiment exposed less difference in problem identification between the four setups. The pairs in field and in lab identified slightly more number of usability problems (51%) than their individual (49%) counterparts. Looking at problem severity, we further found that the individual field sessions identified highest number of all the critical problems namely 9 of the 20 (45%), whereas the individual testers in lab identified 3 out of the 20 critical problems (15%) and the pairs in field and lab experienced 4 each of the 20 critical problems (20%). We found pairs in field identified highest number of serious problem namely 17 out of 58 (29.31%). Similar pattern was found for the serious problems with pairs in lab sessions identifying 16 of 58 problems (27.58%) and individual testers in field found 15 of the 58 problems (25.86%). However, the individual testers in lab found the least number of serious problems namely 10 out of 58 problems (17.24%). Regarding the cosmetic problem identification, we found minor differences between the four setups, which accounts to approximately 25% from each session.

Analyzing the average numbers of identified problems, we found some deviations between the setups; pairs in lab identified 3.44 problems (SD=1.13) also pairs in field identified 3.44 problems (SD=0.53), individual testers in lab identified 3.33 problems (SD=1.50) and in field the individual testers identified 3.22 problems (SD=0.83). The standard deviations indicate variances between the setups and we found no significant differences between the four setups according to a one-way ANOVA test $F(3,32) = 0.091$, $p = 0.965$. Furthermore, we found no significant differences for neither the critical problems $F(3,32) = 1.81$, $p = 0.166$, nor for the identified serious problems $F(3,32) = 1.78$, $p = 0.171$, or for the identified cosmetic problems $F(3,32) = 0.030$, $p = 0.993$.

V. DISCUSSION

The results of the comparative study we performed were surprising, as results did not support much the hypotheses we assumed. The following discussion revisits the research questions set in the introduction:

RQ1: Are same problems found in both lab and field?
According to our study, there was not much difference in the number of problems that were found in four test settings. Our hypothesis that more problems would be found in the field was not supported.

RQ2: Is the severity of the problems same in lab and field?
In terms of problem severity, field testing identified more number of critical and serious problems than the lab testing. However, the number did not vary for the identified cosmetic problems. The hypothesis that more severe problems would be found in the field test was partially supported.

RQ3: Does the test environment affect the user performance?
In the field test, there were interruptions as no restrictions were imposed on the people to move in the field, but these did not seem to affect the performance much. Testers in field were more conscious about the presence of the test evaluator; however, the time taken to complete the tasks was lesser in field testing compared to the lab counterparts. Even though the test as well as the interface was new to them, the children were found to be

relaxed in field. Post task NASA-TLX workload scores revealed that frustration was the least important factor for children when they are tested in the field.

RQ4: Which setting is more suitable to test with the children?

When performing a user interface evaluation with children, even though field-testing may not add significantly to the validity and thoroughness of the test but can be a cost efficient way. Not because of the lesser time taken, but also because the users feel familiar to the place and environment. This outcome supports the results of Hertzum [22], in his study, the time required in the field tests was significantly smaller than in the laboratory test. In the study of Hertzum, the field test was conducted by users, without supervision of the test leader. Based on our study, field tests provided with little more information about the severe problems to improve the user interface and interaction of the system than the lab tests.

VI. CONCLUSION

This study was aimed to examine the effect of location context on the results of usability evaluation with children. Usability evaluation was carried out with 54 children divided to participate in four different settings, two settings in the field and two settings in the usability laboratory. During the test sessions children were required to solve searching tasks on ICDL. Test sessions were recorded. Analyses of the test sessions were done to find the number of usability problems found during each session. The usability problems were classified as cosmetic, serious, and critical based on the severity of the problem found. The results indicate that field evaluation with children uncovered slightly more severe problems than the lab evaluation. The time taken by the children to solve all tasks was lesser in the field. The frustration levels reported were lesser during the field evaluation than during the lab evaluation. Future work would further try to explore the number of usability problem identified by each of the genders separately.

REFERENCES

[1] D.Considine, J. Horton, and G. Moorman, " Teaching and reading the millennial generation through media literacy," Journal of Adolescent & Adult Literacy, 52 (6), 2009, pp. 471–481.

[2] A.Margaryan, A. Littlejohn, and G. Vojt "Are digital natives a myth or reality? University students' use of digital technologies," Computers & Education 56, 2011,pp.429–440.

[3] M.Prensky, " Digital Natives, Digital Immigrants, Part II: Do They Really Think Diffeently?," On the Horizon, Vol.9, No.6, 2001, pp. 15-24.

[4] A.Druin, A. 1999b " The Design of Children's Technology," Morgan Kaufmann Publishers, Inc., San Francisco, CA,1999b

[5] A. Druin, A. and C. Solomon "Designing Multimedia Environments for Children," Wiley & Sons, New York, , 1996.

[6] L. Hanna, K. Risden,, and K.J. Alexander, "Guidelines for Usability Testing with Children," In interactions, September + October 1997, pp. 9 – 14.

[7] T. Tullis, S. Fleischman, M. McNulty, C. Cianchette, and M. Bergel. "An Empirical Comparison of Lab and Remote Usability Testing of Web Sites", Usability Professionals Association Conference, 2002.

[8] A.S. Tsiaousis, G.M. Giaglis, "Evaluating the Effects of the Environmental Context-of-Use on Mobile Website Usability," Mobile Business, 2008. ICMB '08. 7th International Conference on, vol., no., pp.314-322.

[9] K.A. Hummel, A. Hess, and T. Grill, "Environmental context sensing for usability evaluation in mobile hci by means of small wireless sensor networks", In MoMM '08:Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia, pp. 302–306.

[10] M.S. Andreasen, H. Nielsen, S. Schoder, and J. Stage, " What happened to remote usability testing? An empirical study of three methods," Conference on Human Factors in Computing Systems, 2007, pp. 1405-1414.

[11] A. Bruun, P. Gull, L. Hofmeister, and J.Stage, "Let your users do the testing: A comparison of three remote asynchronous usability testing methods," Paper presented at the CHI '09: Proceedings of the 27th International Conference on Human Factors in Computing Systems, Boston, MA, USA.2009, pp.1619-1628.

[12] A. Kaikkonen, T. Kallio, A.Kekäläinen, and A. KankainenCankar, A. "Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing," Journal of Usability studies vol.1, no. 1, 2005, pp 4-16.

[13] F.H.A. Razak, H.Hafit, N. Sedi, N.A. Zubaidi, and H. Haron, (2010) "Usability testing with children: Laboratory vs field studies", User Proc. 2010 International Conference on Science and Engineering (i-USEr), 2010, pp.104-109.

[14] C. Andrrzejczak and D. Liu."The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience," Journal of Systems and Software Vol. 83, Issue 7, 2010.

[15] K.C. Madathil "Synchronous remote usability testing: a new approach facilitated by virtual worlds", Proc. 2011 annual conference on Human factors in computing systems CHI'11, 2011.

[16] L. Baillie, and R.Schatz "Exploring Multimodality in the Laboratory and in the Field," Proc. 7th Intern. Conference on Multimodal Interfaces,2005, pp. 100-107.

[17] C.M. Karat, "Iterative Usability Testing of a Security Application Proceedings of the Human Factors and Ergonomics Society Annual Meeting October 1989 33: pp. 273-277.

[18] R.T., Høegh, J. Kjeldskov, M.B. Skov and J. Stage, "A Field Laboratory for Evaluating in Situ," In Lumsden, J. (Ed.). Handbook of Research on User Interface Design and Evaluation for Mobile Technology, IGI Global, 2008, pp. 982-996.

[19] A. Oztoprak, and C. Erbug, " Field versus Laboratory Usability Testing : A First Comparison, Retrieved from http://www.aydinoztoprak.com/images/HFES_Oztoprak_.pdf

[20] J. Kjeldskov , M.B.Skov, B.S. Als and R.T Høegh "Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field," In Proceedings MobileHCI 2004 conference, Glasgow, UK. Springer- Verlag, 2004. pp.61-73.

[21] R. Molich, Brugervenlige EDB systemer. (Eng: User-Friendly Computer Systems). 2000, Teknisk Forlag.

[22] M. Hertzum, "User Testing in Industry: A case study of laboratory, Workshop and Field Tests," Proceedings of 5th ERCIM WORKSHOP ON "USER INTERFACES FOR ALL , 1999, http://ui4all.ics.forth.gr/UI4ALL-99/Hertzum.pdf